Provable Joint Decontamination for Benchmarking Multiple Large Language Models

Zhenlong Liu[1,2,3], **Hao Zeng**[4], **and Hongxin Wei**[1,*]

[1]Department of Statistics and Data Science, Southern University of Science and Technology
[2]Tencent HY
[3]Shanghai Innovation Institute
[4]School of Mathematical Sciences, University of Electronic Science and Technology of China

Abstract

Benchmark data contamination has become a central challenge in LLM evaluation: when evaluation examples appear in the training data of one or more audited models, reported performance can be inflated and cross-model comparisons become unreliable. A broad line of training-data detection work designs scores to quantify how strongly a model memorizes a given data point, but these score-based methods lack theoretical guarantees. Recent conformal approaches provide provable false-identification control for a *single model*; however, applying them separately to each model can produce model-specific benchmarks, undermining fair comparison across models. In this work, we formalize multi-model benchmark decontamination as a joint selection problem and propose Joint Envelope Conformal Selection (**JECS**), a conformal procedure that enables global contamination rate (GCR) control under stated assumptions. Specifically, JECS computes per-model conformal p-values, aggregates them by the per-item maximum, and reconstructs a conservative envelope of the max-p null distribution from right-tail observations above a data-driven threshold. By applying the adaptive Benjamini–Hochberg (BH) procedure to the envelope-rescaled values, we select a benchmark with provable GCR control. Extensive experiments across various models and benchmarks demonstrate that JECS achieves higher power than the max-p baseline while consistently maintaining the target GCR control. Our code is available at https://github.com/ml-stat-Sustech/Joint-Pure-Benchmark-Selection.

1 Introduction

The remarkable achievements of large language models (LLMs) have been driven in large part by the scale and diversity of their pretraining corpora [Kaplan et al., 2020, Chang et al., 2024]. However, the construction of large pretraining corpora can lead to overlap between training data and evaluation benchmarks, a phenomenon commonly referred to as benchmark data contamination [Sainz et al., 2023, Deng et al., 2024, Xu et al., 2024, Balloccu et al., 2024]. Such overlap can overstate reported performance, obscure whether models truly generalize beyond their training data, and weaken the credibility of benchmark-based evaluation. This highlights the need to construct a decontaminated benchmark for multiple audited models.

To address this issue, a broad line of work [Carlini et al., 2021, Zhang et al., 2025b, Li et al., 2024c] designs detection scores, such as Min-K% [Shi et al., 2024], to quantify how strongly a model memorizes a given data point. Motivated by the lack of theoretical guarantees in these score-based methods, recent work [Liu et al., 2026b] leverages conformal inference to identify training data for

*Correspondence to: Hongxin Wei <weihx@sustech.edu.cn>.

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

a *single model* with provable false-identification control. However, applying such a procedure separately to each model can produce model-specific benchmarks, undermining fair comparison across models. In practical benchmark evaluation, auditors require a shared decontaminated benchmark that supports fair assessment across multiple models. This motivates us to develop methods for selecting *one* shared benchmark with a controlled contamination rate across all audited models.

In this paper, we formulate joint benchmark decontamination as a joint selection problem, where the goal is to select a single benchmark that is decontaminated with respect to multiple audited models. Specifically, we call a sample *jointly pure* if it does not appear in the training data of any audited model, and define the global contamination rate (GCR) as the expected fraction of contaminated samples in the selected set. To control the GCR, we propose Joint Envelope Conformal Selection (**JECS**), which first obtains a valid p-value under the joint null for each candidate by taking the maximum of its per-model conformal p-values. To recover power, we further reconstruct a conservative envelope function for the cumulative distribution function (CDF) of the max-p null distribution, map each max-p value through the fitted function, and then apply the adaptive BH procedure [Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001]. We establish theoretical guarantees showing that JECS controls the contamination rate at a user-specified level α.

Extensive experiments on simulated data and real LLM benchmarks demonstrate the effectiveness of JECS for joint benchmark decontamination. Across all settings, JECS controls the global contamination rate at the user-specified target, while neither the union nor the intersection of per-model conformal selections does so. For instance, on the synthetic setup at $\alpha = 0.1$, JECS attains GCR $= 0.038$, whereas the union and intersection rules violate the target with GCR $= 0.763$ and 0.366, respectively. Furthermore, our method improves power over the max-p baseline. For example, on ArXivTection with Pythia-6.9B [Biderman et al., 2023] at $\alpha = 0.1$ using the Min-K%++ score [Zhang et al., 2025b], our method improves power from 0.094 to 0.447. Overall, these results show that JECS preserves the contamination-rate control required for joint selection while recovering much of the power lost by the conservative max-p baseline.

We summarize our contributions as follows:

1. We formulate multi-model benchmark decontamination as a joint selection problem, where a candidate item is jointly pure only if it is absent from the training data of every audited model. We introduce the global contamination rate (GCR) as the contamination-rate criterion for selecting one shared benchmark across multiple audited models.
2. We propose Joint Envelope Conformal Selection (**JECS**), a joint selection procedure for constructing shared decontaminated benchmarks across multiple audited models. Specifically, JECS mitigates the super-uniform conservatism of the valid max-p baseline, thereby retaining global contamination rate control while substantially improving power.
3. We establish theoretical guarantees showing that JECS controls GCR under stated conditions. We further validate these guarantees through extensive experiments, demonstrating the effectiveness of JECS on both simulated data and real-world LLM benchmarks.

2 Preliminary

Setup. Let θ be a target language model trained on a private corpus $\mathcal{D}_{\text{train}}(\theta)$, and let $\mathcal{X}$ denote the input space of token sequences. For a sample $x \in \mathcal{X}$, the binary membership indicator

$$M(x;\theta) = \mathbb{1}\{x \in \mathcal{D}_{\text{train}}(\theta)\} \in \{0,1\} \tag{1}$$

records whether x was used to train θ. Specifically, $M = 1$ indicates a member and $M = 0$ a non-member. Throughout the paper, we audit K such models $\{\theta_k\}_{k=1}^K$ on a candidate pool $\{x_i\}_{i=1}^n$, and write $M_i^k = M(x_i;\theta_k)$ for the membership status of candidate instance i under model θ_k.

Training data detection. Training data detection, also termed membership inference attack (MIA), aims to infer the membership indicator $M(x;\theta)$ from query access to the model p_θ. The standard pipeline assigns each sample a detection score $T(x;\theta)$ and predicts membership by a level-set rule, with the inequality direction determined by the score convention. In this paper, we orient T so that smaller values provide stronger evidence of non-membership, and hence a threshold rule may be written as $\widehat{M} = \mathbb{1}\{T(x;\theta) \geq \tau\}$ with a threshold τ determined by a validation set [Shokri

et al., 2017, Ye et al., 2022]. A representative score is Min-K%++ [Zhang et al., 2025b], which for a sequence $x = (t_1, \ldots, t_L)$ averages a normalized log-probability over the bottom-$K\%$ tokens:

$$T_{\text{MIN-K\%++}}(x;\theta) = \frac{1}{|\mathcal{I}|}\sum_{\ell\in\mathcal{I}} \frac{\log p_\theta(t_\ell \mid t_{<\ell}) - \mu_\ell}{\sigma_\ell}, \qquad \mu_\ell = \mathbb{E}_{t\sim p_\theta(\cdot|t_{<\ell})}[\log p_\theta(t \mid t_{<\ell})], \tag{2}$$

where σ_ℓ is the standard deviation of $\log p_\theta(\cdot \mid t_{<\ell})$ under the next-token distribution and $\mathcal{I} \subset \{1, \ldots, L\}$ indexes the lowest $K\%$ tokens by per-token log-probability. Other widely used scores include perplexity [Carlini et al., 2021], Min-K% [Shi et al., 2024], and Modified Entropy [Song and Mittal, 2021]. These scores provide useful per-instance evidence of memorization by a specific model, but they do not provide theoretical guarantees for the resulting predictions.

Conformal inference for training data identification. To provide provable evidence, recent work [Liu et al., 2026b] formulates training-data identification as a multiple-hypothesis testing problem. For each audited model θ_k, the model-wise null and alternative are:

$$H^k_{0,i} : x_i \in \mathcal{D}_{\text{train}}(\theta_k), \qquad H^k_{1,i} : x_i \notin \mathcal{D}_{\text{train}}(\theta_k). \tag{3}$$

Rejecting $H^k_{0,i}$ means that x_i is identified as pure with respect to θ_k; a false identification occurs when a contaminated item is incorrectly selected.[1] The conformal p-value [Vovk et al., 2003, 2005] calibrates the detection scores into valid p-values:

$$p^k_i = \frac{1 + \sum_{x_\ell\in\mathcal{D}_{\text{cal}}} \mathbf{1}\{T(x_\ell;\theta_k) \le T(x_i;\theta_k)\}}{|\mathcal{D}_{\text{cal}}| + 1}, \tag{4}$$

where $\mathcal{D}_{\text{cal}}$ is a calibration set of known members of θ_k drawn from the same covariate distribution as contaminated candidates. The Benjamini–Hochberg (BH) procedure [Benjamini and Hochberg, 1995] is then used to select a subset:

$$\mathcal{S}_k = \left\{ i : p^k_i \le \frac{\alpha\, r^*}{n} \right\}, \text{ where } r^* = \max\left\{ r \in \{1, \ldots, n\} : p^k_{(r)} \le \frac{\alpha\, r}{n} \right\}, \tag{5}$$

where $p^k_{(1)} \le \cdots \le p^k_{(n)}$ are the sorted model-wise p-values. In the original training-data-identification setting, prior works [Bates et al., 2023, Liu et al., 2026b] have shown that the expected fraction of θ_k-non-members among the items identified as members can be controlled at a user-specified level $\alpha \in (0, 1)$. However, this guarantee is model-specific: applying the procedure separately to different audited models yields different selected subsets $\mathcal{S}_1, \ldots, \mathcal{S}_K$. In practical benchmark evaluation, auditors need a single decontaminated benchmark for a fair comparison across all audited models. This gap motivates the joint benchmark decontamination problem below.

3 Joint Benchmark Decontamination

3.1 Problem formulation

We formulate the **joint benchmark decontamination** problem, whose goal is to select one shared benchmark that is decontaminated with respect to all audited models. Unlike the model-wise setting above, contamination in this problem is defined globally across the audited model collection. An instance x_i is *contaminated* if it appears in the training data of at least one audited model, and is *jointly pure* only if it appears in none. Accordingly, the joint null and alternative hypotheses are:

$$H_{0,i} : \exists k^* \in \{1, \ldots, K\},\ x_i \in \mathcal{D}_{\text{train}}(\theta_{k^*}), \qquad H_{1,i} : \forall k,\ x_i \notin \mathcal{D}_{\text{train}}(\theta_k). \tag{6}$$

Rejecting $H_{0,i}$ indicates that x_i is inferred to be jointly pure and is therefore selected for inclusion in the benchmark. We therefore quantify error by the contamination rate within the selected benchmark. Formally, for a selected set $\mathcal{S} \subseteq \{1, \ldots, n\}$, we define the global contamination proportion (GCP) and global contamination rate (**GCR**) as follows:

$$\text{GCP}(\mathcal{S}) = \frac{\sum_{i=1}^n \mathbb{1}\{i \in \mathcal{S},\ H_{0,i} \text{ is true}\}}{1 \vee |\mathcal{S}|}, \qquad \text{GCR}(\mathcal{S}) = \mathbb{E}[\text{GCP}(\mathcal{S})], \tag{7}$$

[1] The testing direction here is reversed from the convention in Liu et al. [2026b], where the null is non-member. We reverse the null because benchmark construction selects non-members.

where we denote $a \vee b = \max\{a, b\}$. Our goal is to return a subset $\mathcal{S}$ such that $\mathrm{GCR}(\mathcal{S}) \leq \alpha$, where $\alpha \in (0, 1)$ is a user-specified level. Beyond GCR control, we also seek a procedure that selects as many true candidates as possible. To quantify this objective, we define the **power** as

$$\mathrm{Power}(\mathcal{S}) = \mathbb{E}\left[\frac{\sum_{i=1}^{n} \mathbb{1}\{i \in \mathcal{S}, H_{1,i} \text{ is true}\}}{1 \vee \sum_{i=1}^{n} \mathbb{1}\{H_{1,i} \text{ is true}\}}\right]. \tag{8}$$

An ideal method should control the GCR at the prescribed level α with power as high as possible.

Naive per-model composition fails to control GCR. A natural but invalid approach is to first form the model-wise selections $\mathcal{S}_1, \ldots, \mathcal{S}_K$ from the conformal p-values in Equation (4) using the single-model rule in Equation (5), and then combine them by union or intersection. The union $\cup_k \mathcal{S}_k$ is too liberal: an item may be selected because it appears pure for one model while being contaminated for another. The intersection $\cap_k \mathcal{S}_k$ is more conservative, but still lacks GCR control because the denominator changes after composition and the model-wise selections are dependent. As shown in Table 1, both composition violate the target on a controlled synthetic setup with $K = 4$ audited models and known joint-purity labels (full configuration in Appendix E.1): Union stays near GCP ≈ 0.76, and Intersection increasingly exceeds α as α grows. Thus, joint GCR control for Equation (6) requires testing $H_{0,i}$ directly through an aggregated statistic rather than composing per-model decisions.

Table 1: Realized GCP of naive per-model composition on synthetic data. Entries marked with a † exceed α, indicating failure to meet the GCR target.

Procedure	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
Union $\cup_k \mathcal{S}_k$	$0.763^\dagger$	$0.764^\dagger$	$0.764^\dagger$
Intersection $\cap_k \mathcal{S}_k$	$0.366^\dagger$	$0.614^\dagger$	$0.757^\dagger$
JECS	0.038	0.087	0.162

3.2 A max-p baseline: Joint Max Conformal Selection (JMCS)

We now adapt the model-wise conformal p-values from Section 2 to the joint selection problem. Throughout this section, we use a shared calibration set $\mathcal{D}_{\mathrm{cal}} = \{x_\ell\}_{\ell=n+1}^{n+m}$ whose items are known members of every audited model, i.e., $x_\ell \in \mathcal{D}_{\mathrm{train}}(\theta_k)$ for all $k \in \{1, \ldots, K\}$, and are drawn from the same candidate-pool covariate distribution. In practice, constructing $\mathcal{D}_{\mathrm{cal}}$ only requires high-confidence shared members rather than each model's full training corpus. As discussed in Appendix F, such items can be collected from public benchmark training splits routinely mixed into LLM training, provider-disclosed pretraining sources such as Common Crawl, Wikipedia, arXiv, and GitHub, and regulatory training-data disclosures, then restricted to the benchmark domain and matched on score-relevant properties including topic, programming language, and sequence length.

Applying the conformal construction in Equation (4) to each audited model yields model-wise p-values $\{p_i^k\}_{k=1}^K$. By exchangeability, for every model-wise null $H_{0,i}^k$,

$$\mathbb{P}(p_i^k \leq t \mid H_{0,i}^k) \leq t, \qquad t \in [0, 1]. \tag{9}$$

To test the joint null $H_{0,i}$ via one statistic, we aggregate the K model-wise conformal p-values by their maximum:

$$p_i^* = \max_{1 \leq k \leq K} p_i^k. \tag{10}$$

The aggregated value p_i^* is small only if every model's p-value is small, which is the operational meaning of joint purity. The following lemma states that p_i^* is itself a valid p-value for the joint null.

Lemma 1 (Validity of the max-p)**.** *If Equation (9) holds for each model-wise null, then p_i^* is valid for the joint null:*

$$\mathbb{P}(p_i^* \leq t \mid H_{0,i}) \leq t, \qquad t \in [0, 1]. \tag{11}$$

Proof. Under $H_{0,i}$, there exists k^* such that $H_{0,i}^{k^*}$ is true. Since $p_i^* = \max_k p_i^k \geq p_i^{k^*}$, we have $\mathbb{P}(p_i^* \leq t \mid H_{0,i}) \leq \mathbb{P}(p_i^{k^*} \leq t \mid H_{0,i}^{k^*}) \leq t$. □

Lemma 1 immediately delivers a baseline procedure: apply the Benjamini–Hochberg (BH) [Benjamini and Hochberg, 1995] step-up rule to the maxima $\{p_i^*\}_{i=1}^n$ at level α. Letting $p_{(1)}^* \leq \cdots \leq$

Algorithm 1 Joint Envelope Conformal Selection (JECS)

Require: Candidate items $\{x_i\}_{i=1}^n$, audited models $\{\theta_k\}_{k=1}^K$, calibration data $\mathcal{D}_{\text{cal}} = \{x_i\}_{i=n+1}^{n+m}$, target level α, tail threshold λ.

1: **for** $k = 1, \dots, K$ **do**
2: Construct conformal p-values $\{p_i^k\}_{i=1}^n$ via Equation (4).
3: **end for**
4: Aggregate $p_i^* = \max_k p_i^k$ for $i = 1, \dots, n$.
5: Estimate $\widehat{g}(\lambda^+)$ and $\widehat{G}_n$ from the right tail $\{p_i^* : p_i^* > \lambda\}$.
6: Form the envelope $\widehat{F}_{\text{fit}}$ via Equation (14) and Equation (15).
7: Transform $\widetilde{p}_i = \widehat{F}_{\text{fit}}(p_i^*)$ and estimate $\widehat{\pi}_0$ via Equation (16).
8: Return the selected set $\mathcal{S}$ via Equation (17).

$p_{(n)}^*$ be the sorted maxima, the rule selects

$$\mathcal{S} = \left\{ i : p_i^* \le \frac{\alpha\, r^*}{n} \right\}, \text{ where } r^* = \max\left\{ r \in \{1, \dots, n\} : p_{(r)}^* \le \frac{\alpha\, r}{n} \right\}, \tag{12}$$

We refer to this procedure as the Joint Max Conformal Selection (JMCS). Lemma 1 together with the standard BH analysis [Benjamini and Hochberg, 1995] yields finite-sample GCR control under the joint null. JMCS is the natural minimum-viable procedure for joint pure benchmark selection, and to our knowledge no prior work has analyzed it as such; we use it as the baseline against which JECS is compared in Section 4, and summarize the full procedure in Algorithm 2 of Appendix C.

The super-uniformity tax. JMCS is valid in finite samples but substantially conservative. The BH cutoff $\alpha r/n$ is calibrated to the uniform scale, whereas under $H_{0,i}$ the maximum statistic $p_i^* = \max_k p_i^k$ is strictly super-uniform: in the stylized homogeneous null with $p_i^1, \dots, p_i^K \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, the null CDF is $F_0(t) = t^K$ and the density $f_0(t) = Kt^{K-1}$ vanishes at $t = 0$ and concentrates near $t = 1$ (Figure 1, $K = 8$). Almost all null mass therefore sits above the BH cutoff: a rejection requires $p_i^* \le \alpha r/n$, far below the natural null scale $(r/n)^{1/K}$, and the gap widens as α tightens or K grows. Although Lemma 1 guarantees GCR control, this procedure can lose substantial power as K increases.

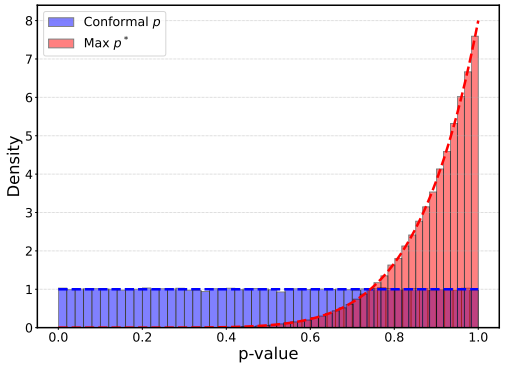


Figure 1: Null density of the per-model conformal p versus $p_i^* = \max_k p_i^k$ at $K = 8$.

3.3 Joint Envelope Conformal Selection (JECS)

As discussed above, p_i^* is expected to follow a monotone increasing distribution on $[0, 1]$ under the null, resembling a Beta distribution and concentrating near 1. While this preserves validity, it often leads to overly conservative selection. A natural remedy is therefore to estimate the null CDF of p_i^* and use the estimated CDF to recalibrate p_i^*, improving selection power while retaining validity.

Let F_0 and F_1 denote the CDFs of p_i^* under contaminated and pure items, and let $F_{\text{mix}} = \pi_0 F_0 + (1 - \pi_0) F_1$ be the marginal mixture. The empirical asymmetry that drives JECS is that pure items concentrate near the left tail of p_i^*, while the right tail $\{p_i^* > \lambda\}$ is enriched for contaminated items. We exploit this asymmetry by anchoring a conservative envelope to the right tail and propagating it leftward. For a fixed threshold $\lambda \in (0, 1)$, define the right-tail conditional CDF

$$G(x) = \mathbb{P}(P^* \le x \mid P^* > \lambda) = \frac{F(x) - F(\lambda)}{1 - F(\lambda)}, \qquad x \in [\lambda, 1], \tag{13}$$

where P^* denotes a generic draw from the distribution with CDF F. Let $g(\lambda^+)$ denote the corresponding right-boundary density. From the observed maxima $\{p_i^*\}_{i=1}^n$, JECS estimates the right-

tail mixture CDF $\widehat{G}_n$ on $[\lambda, 1]$ and the boundary density $\widehat{g}(\lambda^+)$, both implemented by a k-nearest-neighbor estimator. The fitted envelope at the anchor is

$$\widehat{F}_{\text{fit}}(\lambda) = \frac{\lambda\,\widehat{g}(\lambda^+)}{1 + \lambda\,\widehat{g}(\lambda^+)}, \tag{14}$$

and the global envelope is the piecewise reconstruction

$$\widehat{F}_{\text{fit}}(x) = \begin{cases} \dfrac{\widehat{F}_{\text{fit}}(\lambda)}{\lambda}\,x, & x \in (0, \lambda], \\ \widehat{F}_{\text{fit}}(\lambda) + \big(1 - \widehat{F}_{\text{fit}}(\lambda)\big)\widehat{G}_n(x), & x \in (\lambda, 1]. \end{cases} \tag{15}$$

The left branch is linear with slope $\widehat{F}_{\text{fit}}(\lambda)/\lambda$, whereas the right branch uses the empirical right-tail CDF rescaled to match the anchor. Figure 2 illustrates the role of this fitted envelope: it maps the raw max-p values to a conservative CDF scale before BH-style selection. The threshold λ marks the point beyond which the right tail is dominated by contaminated items; its data-driven choice and sensitivity are deferred to Section 4. Each maximum-p value is then transformed through the envelope as $\widetilde{p}_i = \widehat{F}_{\text{fit}}(p_i^*)$.

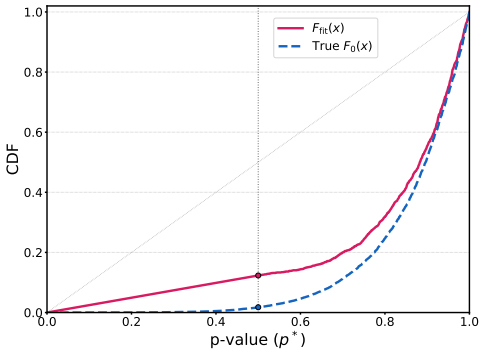


Figure 2: Fitted envelope schematic. The fitted $\widehat{F}_{\text{fit}}$ is a conservative estimate of the true F_0.

To improve power, we use Storey's estimator [Storey, 2002] on the transformed values $\{\widetilde{p}_i\}$, using the null-proportion estimate

$$\widehat{\pi}_0 = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{p_i^* > \lambda\}}{1 - \widehat{F}_{\text{fit}}(\lambda)}, \tag{16}$$

and, with $\widetilde{p}_{(1)} \le \cdots \le \widetilde{p}_{(n)}$ the sorted transformed values, returning

$$\mathcal{S} = \left\{ i : \widetilde{p}_i \le \frac{\alpha\, r^*}{\widehat{\pi}_0\, n} \right\}, \text{ where } r^* = \max\left\{ r : \widetilde{p}_{(r)} \le \frac{\alpha\, r}{\widehat{\pi}_0\, n} \right\}, \tag{17}$$

with $r^* = 0$ and $\mathcal{S} = \emptyset$ when the set is empty. Relative to JMCS, JECS replaces p_i^* with $\widetilde{p}_i$ to mitigate the conservatism induced by super-uniformity, and replaces the fixed choice $\widehat{\pi}_0 = 1$ with the data-adaptive estimate in Equation (16). To establish the theoretical guarantee, we first state the key property motivating the construction of $\widehat{F}_{\text{fit}}(p_i^*)$:

Proposition 2 (Envelope domination)**.** *Let F_{fit} be the population counterpart of Equation (15). Then $F_{\text{fit}}(x) \ge F_0(x)$ for all $x \in (0, 1]$.*

Proposition 2 shows that F_{fit} dominates F_0 pointwise. Consequently, the transformed values $\widetilde{p}_i$ remain super-uniform under the joint null, and the estimator $\widehat{\pi}_0$ in Equation (16) is conservative for the true null proportion. Combined with the standard Storey-BH argument, these two facts yield asymptotic GCR control, as formalized below:

Theorem 3 (Asymptotic GCR control)**.** *Under the assumptions stated in Appendix B.1, the selected set returned by JECS satisfies*

$$\limsup_{n \to \infty} \ \mathrm{GCR}(\mathcal{S}) \le \alpha. \tag{18}$$

The proofs of Proposition 2 and Theorem 3 are deferred to Appendices B.2 and B.3. This guarantee relies on the validity of the model-wise conformal p-values under exchangeability between shared-member calibration items and contaminated candidates, the convexity of the null max-p CDF in Assumption 4, and the right-tail distinguishability condition in Assumption 5 under which larger max-p values reflect stronger evidence of contamination.

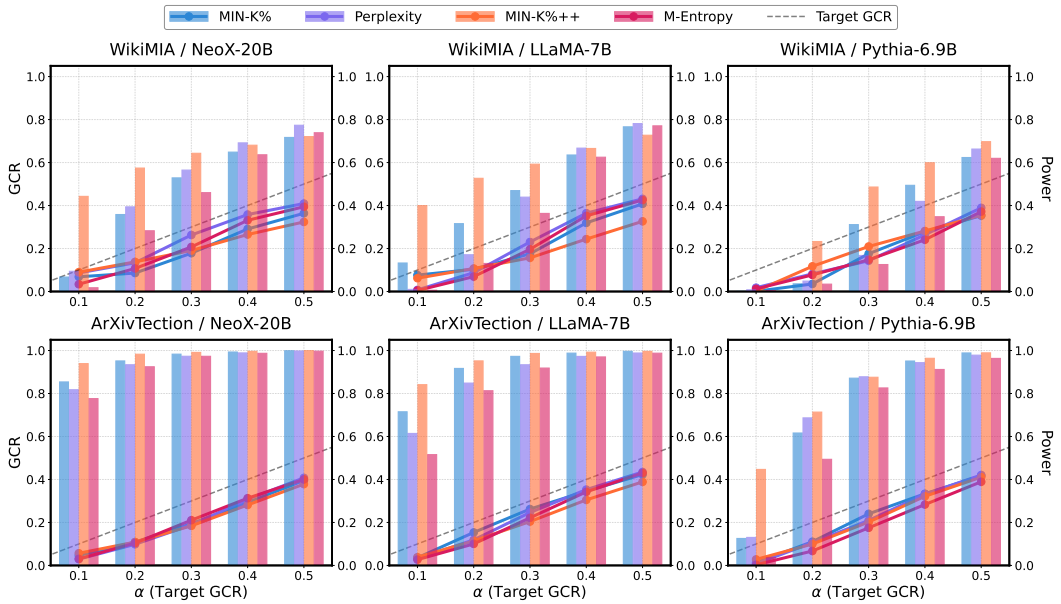


Figure 3: Joint GCR control with JECS at $K = 16$. Each panel corresponds to a dataset–model family pair and reports realized GCP curves on the left axis and Power bars on the right axis for four detection scores under JECS; the dashed diagonal is the target GCR.

4 Experimental results

In this section, we empirically evaluate JECS on training-data-detection benchmarks with a controlled member/non-member structure, and verify its joint GCR control as well as its power against the JMCS baseline of Section 3.2. Throughout, JECS is instantiated with the data-driven right-tail threshold rule and the Storey null-proportion estimator of Equation (16), which we adopt as the default configuration.

4.1 Setup

Datasets and audited models. The main evaluation uses WikiMIA [Shi et al., 2024] and ArXivTection [Duarte et al., 2024] with $K = 16$ independently fine-tuned models initialized from GPT-NeoX-20B [Black et al., 2022], Pythia-6.9B [Biderman et al., 2023], and LLaMA-7B [Touvron et al., 2023]; appendix experiments additionally cover five MIMIR [Duan et al., 2024] subsets (HackerNews, DMMath, GitHub, Pile-CC, PubMed) in Appendix E.2 and a mixed-family audit pool spanning Pythia (1.4B/2.8B/6.9B/12B), GPT-NeoX-20B, and LLaMA (7B/13B/30B) in Appendix E.5. The contamination-controlled three-block split, the resulting per-instance membership labels M_i^k, the contamination fraction ρ, and per-dataset sample counts are deferred to Appendix D.2.

Detection scores and baseline. We employ the per-model score $T(x; \theta_k)$ with four standard detectors: Perplexity [Carlini et al., 2021], Min-K% [Shi et al., 2024], Min-K%++ [Zhang et al., 2025b], and Modified Entropy [Song and Mittal, 2021].

4.2 Main results

JECS controls GCR across detection scores and model families. We next evaluate JECS on the two real benchmarks across all three model families and four detection scores. Figure 3 reports realized GCP curves (left axis) and Power bars (right axis) over the various levels α at $K = 16$. Throughout these evaluations, the realized GCP stays at or below the target diagonal. The procedure thus delivers controlled joint pure subsets at the user-specified level without configuration-specific tuning, and the control is stable across detection scores. The same control pattern is verified on five additional MIMIR subsets in Appendix E.2, and on a heterogeneous mixed-family audit pool of Pythia, GPT-NeoX, and LLaMA variants in Appendix E.5.

Table 2: Power on ArXivTection at $K = 16$. Reported values are average Power with standard error. "Base" refers to JMCS and "Ours" refers to JECS. **Bold** numbers mark the better result in each pair.

Model	Score	$\alpha = 0.1$		$\alpha = 0.2$		$\alpha = 0.3$	
		Base	Ours	Base	Ours	Base	Ours
NeoX-20B	Perplexity	0.403±0.008	**0.817**±0.002	0.829±0.001	**0.934**±0.001	0.909±0.000	**0.973**±0.000
	Min-K%++	0.872±0.001	**0.939**±0.001	0.944±0.000	**0.983**±0.000	0.972±0.000	**0.991**±0.000
	Min-K%	0.576±0.003	**0.854**±0.002	0.871±0.001	**0.951**±0.001	0.926±0.000	**0.983**±0.000
	M-Entropy	0.235±0.008	**0.776**±0.002	0.783±0.001	**0.924**±0.001	0.886±0.001	**0.973**±0.001
LLaMA-7B	Perplexity	0.040±0.004	**0.614**±0.005	0.579±0.002	**0.848**±0.002	0.755±0.001	**0.934**±0.002
	Min-K%++	0.715±0.001	**0.841**±0.002	0.844±0.001	**0.952**±0.001	0.918±0.000	**0.986**±0.000
	Min-K%	0.055±0.004	**0.715**±0.003	0.680±0.001	**0.917**±0.002	0.793±0.001	**0.973**±0.001
	M-Entropy	0.026±0.003	**0.516**±0.005	0.523±0.002	**0.813**±0.002	0.716±0.001	**0.918**±0.001
Pythia-6.9B	Perplexity	0.007±0.001	**0.130**±0.006	0.149±0.004	**0.687**±0.006	0.464±0.004	**0.878**±0.002
	Min-K%++	0.094±0.004	**0.447**±0.002	0.465±0.001	**0.713**±0.002	0.627±0.001	**0.876**±0.002
	Min-K%	0.006±0.001	**0.126**±0.007	0.139±0.005	**0.616**±0.005	0.460±0.002	**0.871**±0.002
	M-Entropy	0.004±0.001	**0.037**±0.004	0.095±0.004	**0.494**±0.007	0.370±0.003	**0.826**±0.002

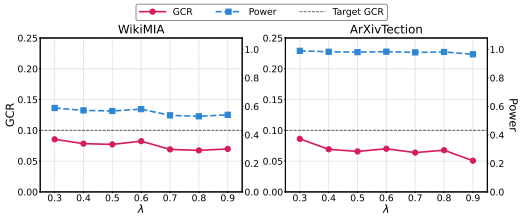


Figure 4: GCR and Power curves over λ with $\alpha = 0.1$. The dashed line is the target GCR.

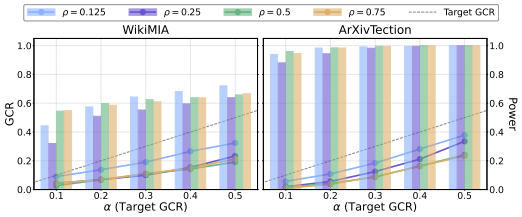


Figure 5: GCP curve (left axis) and Power bars (right axis) under varying training fraction ρ.

JECS attains higher Power than JMCS. Table 2 reports the Power comparison between JECS and JMCS on ArXivTection at $K = 16$. JECS consistently achieves higher Power than JMCS across target levels, model families, and detection scores, with the largest gains at smaller α. For example, under LLaMA-7B at $\alpha = 0.1$ using Min-K%, JECS improves Power from 0.055 to 0.715 over JMCS. The corresponding results on WikiMIA are provided in Appendix E.3.

4.3 Sensitivity analysis

Effect of the right-tail threshold λ. Figure 4 sweeps the right-tail threshold $\lambda \in \{0.3, \ldots, 0.9\}$ on WikiMIA and ArXivTection (NeoX-20B, $K = 8$, Min-K%++, $\alpha = 0.1$) and overlays the data-driven rule. The adaptive choice tracks the best fixed λ in power without producing any GCP violation. This is also the empirical check for the right-tail dominance condition behind Theorem 3: even at the smallest λ tested, GCP stays below α, indicating that the right tail is contaminated-dominated across the operating range.

Impact of the number of audited models K on control and power. Figure 6 varies $K \in \{2, 4, 8, 16\}$ on WikiMIA across all three model families with Min-K%++. GCP stays under the target diagonal as K grows, while Power does not collapse with K. Our claim is not that every Power curve is strictly monotone in K, but that the procedure remains controlled as the number of jointly audited models grows; individual power trends can be non-monotone because, by Equation (47), the joint-purity prevalence in the test mixture depends on K through the term $(1 - \rho)^K$ from the remainder block while the positives in the globally pure block stay fixed.

Robustness of error control to variations in the training fraction ρ. The training fraction ρ controls the contamination level in our experiments: it is the per-model sampling fraction within the remainder block of the controlled three-block split, as defined in Appendix D.2. Figure 5 sweeps

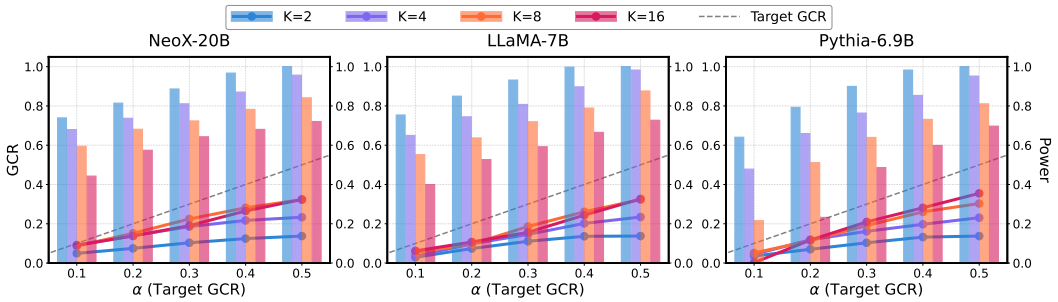


Figure 6: Sensitivity to the number of audited models K on WikiMIA. GCP (lines, left axis) and Power (bars, right axis) with Min-K%++.

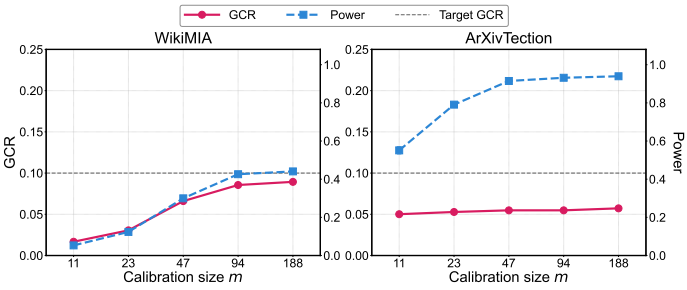


Figure 7: GCR and Power curves over the calibration size m with $\alpha = 0.1$. The dashed line is the target GCR.

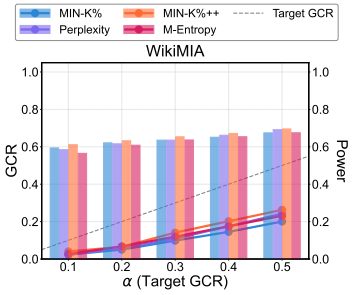


Figure 8: GCP curves (left axis) and Power bars (right axis) on an eight-family audit pool.

$\rho \in \{0.125, 0.25, 0.5, 0.75\}$ on NeoX-20B at $K = 16$ with Min-K%++, on both WikiMIA and ArXivTection. Across all four values of ρ and both datasets, the realized GCP stays at or below the target diagonal, confirming that JECS's control is robust to the contamination level.

Effect of the calibration size m. Figure 7 varies the number of shared calibration items $m \in \{11, 23, 47, 94, 188\}$ on WikiMIA and ArXivTection with NeoX-20B at $K = 16$, Min-K%++, and $\alpha = 0.1$, where smaller calibration sets are nested subsets of larger ones. The realized GCP stays below α at every calibration size on both datasets. Power increases with m and largely saturates beyond $m = 94$: doubling the calibration set to $m = 188$ raises power by only 1.5 and 0.8 percentage points on WikiMIA and ArXivTection, respectively. A calibration set of around one hundred shared members is therefore sufficient for JECS to reach its full power in these settings.

Robustness to heterogeneous audited model families. To test whether JECS requires the audited models to share an architecture, we construct a $K = 16$ audit pool on WikiMIA from eight open-weight families, namely Qwen2.5-7B, Mistral-7B, OPT-6.7B, Pythia-6.9B [Biderman et al., 2023], GPT-NeoX-20B [Black et al., 2022], LLaMA-7B [Touvron et al., 2023], Gemma-2-9B, and OLMo-2-7B, each contributing two adapters fine-tuned on distinct model-specific subsets. As shown in Figure 8, the realized GCP stays below the target diagonal for all four detection scores across the evaluated levels, and Power at $\alpha = 0.1$ ranges from 0.57 to 0.61. The control of JECS therefore carries over to audit pools that mix architectures and pretraining corpora.

5 Related Work

Benchmark data contamination. Benchmark data contamination occurs when evaluation examples overlap with a model's training data, which can inflate reported performance [Sainz et al., 2023, Balloccu et al., 2024, Xu et al., 2024, Li, 2024, Bordt et al., 2025, Cheng et al., 2025]. One line of work reduces this risk before evaluation through dynamic benchmark construction [Zhu et al., 2024a, Li et al., 2024a, White et al., 2025, Srivastava et al., 2026], data rewriting [Yang et al., 2023, Zhu et al., 2024b], or contamination-free benchmark curation [Zhao et al., 2024]. Another line detects

contamination after models and benchmarks already exist. Most methods in this setting operate at the sample level, inferring whether a particular example was used for training [Shokri et al., 2017, Carlini et al., 2022, Liu et al., 2026a,b]; such methods have been studied for LLMs [Carlini et al., 2021, Shi et al., 2024, Zhang et al., 2024b, Golchin and Surdeanu, 2024, Li and Flanigan, 2024, Zhang et al., 2025a, Hu et al., 2025b, Zhang et al., 2025b, Mattern et al., 2023, Xie et al., 2024, Zhang et al., 2024b, Raoof et al., 2025, Kaneko et al., 2025, Yi and Li, 2026], vision language models [Ko et al., 2023, Li et al., 2024c, Hu et al., 2025a, LIU et al., 2026]. Dataset-level detection methods instead test for contamination at a broader dataset scale [Vu et al., 2023, Oren et al., 2024, Ko et al., 2023, Li et al., 2024b, Maini et al., 2024, Zhang et al., 2024a, Choi et al., 2025, Zawalski et al., 2026]. Both lines of work target a single audited model or a single dataset and do not return a shared item-level decontaminated subset across multiple audited models. In contrast, JECS aggregates per-model instance-level conformal p-values into a joint selection procedure that returns one benchmark shared by K audited models with provable GCR control.

Conformal selection. Selective inference with false discovery rate control is rooted in the Benjamini–Hochberg (BH) procedure on valid p-values [Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001, Storey et al., 2004]. Conformal prediction provides a distribution-free route to such p-values under exchangeability [Vovk et al., 2005, Papadopoulos, 2008], and recent work combines conformal calibration with BH-type selection to obtain finite-sample selection guarantees [Bates et al., 2023, Jin and Candès, 2023]. Subsequent extensions address more constrained settings, including covariate shift through weighted conformal scores [Jin and Candès, 2025], multivariate responses through vector-valued conformity scores [Bai et al., 2025b], and multiple response conditions through conditional conformity constructions [Hao et al., 2026]. These ideas have also been applied to LLM alignment [Gui et al., 2024], drug discovery [Bai et al., 2025a], and training-data identification [Liu et al., 2026b]. The closest formal comparison is multivariate conformal selection [Bai et al., 2025b], where the vector structure comes from multiple response coordinates within one prediction problem. In joint benchmark decontamination, the multiplicity comes from audited models: each item has K membership statuses, and it is contaminated if any audited model trained on it. This yields a joint conformal selection problem under a union null: the procedure must aggregate dependent model-wise conformal evidence into one item-level ranking while controlling the expected fraction of union-null items among selected items.

6 Conclusion

In this paper, we formalize joint benchmark decontamination as a joint selection problem, where the goal is to select one shared benchmark whose global contamination rate is controlled across all audited models. We propose Joint Envelope Conformal Selection (JECS), which computes model-wise conformal p-values, aggregates them through a max-p statistic, reconstructs a conservative envelope for the max-p null distribution, and applies adaptive Storey-BH to the transformed values. Experiments on synthetic data and LLM benchmarks show that JECS maintains empirical contamination control across datasets, model families, and detection scores, while improving power over JMCS. The method is agnostic to the choice of detection score and applies beyond LLM evaluation whenever shared-member calibration data are available.

Limitations. The current procedure relies on a calibration set whose distribution is close to the candidate pool, which is a standard condition for conformal calibration. While importance-weighted conformal p-values [Jin and Candès, 2025] can be incorporated into JECS to handle covariate shift along observed features, establishing formal guarantees under broader distribution shifts remains an important direction for future work. In addition, our formulation focuses on exact instance-level membership, and extending the framework to soft contamination such as near-duplicates, paraphrases, or translated variants is another promising direction.

Acknowledgments and Disclosure of Funding

This research is supported by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515011367) . This project is also supported by the Jiangsu Provincial Key Discipline Construction Project (Statistics) and the open project of the Joint Lab for Statistics and Finance (Grant No. 2025JLSF101). Hao Zeng is supported by the SUSTech–NUS Joint Research Program. We

thank Shenzhen Guangming Science City Development and Construction Co., Ltd. and its Material Genome Big-Science Facilities Platform for providing technical support and assistance in data collection and analysis.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Tian Bai, Peng Tang, Yuting Xu, Vladimir Svetnik, Bingjia Yang, Abbas Khalili, Xiang Yu, and Archer Y Yang. Conformal selection for efficient and accurate compound screening in drug discovery. *Journal of Chemical Information and Modeling*, 65(24):13070–13085, 2025a.

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

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, 2024.

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

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

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

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

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé, editors, *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

Sebastian Bordt, Suraj Srinivas, Valentyn Boreiko, and Ulrike von Luxburg. How much can we forget about data contamination? In *ICML*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

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

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

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

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

Hyeong Kyu Choi, Maxim Khanov, Hongxin Wei, and Yixuan Li. How contaminated is your benchmark? measuring dataset leakage in large language models with kernel divergence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wVDR2qmE28.

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

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

André Vicente Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. DE-COP: Detecting copyrighted content in language models training data. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=LO4xhXmFal.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, OJ L, 2024/1689, 12 July 2024, July 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj. CELEX: 32024R1689.

Christian Genest, Étienne Marceau, and Mhamed Mesfioui. Compound poisson approximations for individual models with dependent risks. *Insurance: Mathematics and Economics*, 32(1):73–91, 2003.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37:73884–73919, 2024.

Qingyang Hao, Wenbo Liao, Bingyi Jing, and Hongxin Wei. Multi-condition conformal selection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=giL8Q1V26J.

Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. Membership inference attacks against vision-language models. In *USENIX Security Symposium*, pages 1589–1608. USENIX Association, 2025a.

Zirui Hu, Yingjie Wang, Zheng Zhang, Hong Chen, and Dacheng Tao. A statistical approach for controlled training data detection. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=XAN8G0rvoB.

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

Ying Jin and Emmanuel J Candès. Model-free selective inference under covariate shift via weighted conformal p-values. *Biometrika*, page asaf066, 2025.

Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. Sampling-based pseudo-likelihood for membership inference attacks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8894–8907, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.465. URL https://aclanthology.org/2025.findings-acl.465/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881, 2023.

Takaaki Koike, Liyuan Lin, and Ruodu Wang. Invariant correlation under marginal transforms. *Journal of Multivariate Analysis*, 204:105361, 2024.

Changmao Li and Jeffrey Flanigan. Task contamination: language models may not be few-shot anymore. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i16.29808. URL https://doi.org/10.1609/aaai.v38i16.29808.

Yanyang Li. Awesome data contamination. https://github.com/lyy1994/awesome-data-contamination, 2024.

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

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. An open-source data contamination report for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.30. URL https://aclanthology.org/2024.findings-emnlp.30/.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c. URL https://openreview.net/forum?id=nv2Qt5cj1a.

Yihao LIU, Xinqi LYU, Dong Wang, Yanjie Li, and Bin Xiao. LOMIA: Label-only membership inference attacks against pre-trained large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=7JjS2cdBYN.

Zhenlong Liu, Wenyu Jiang, Feng Zhou, and Hongxin Wei. How does bayesian sampling help membership inference attacks? In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=rSytT800kl.

Zhenlong Liu, Hao Zeng, Weiran Huang, and Hongxin Wei. Provable training data identification for large language models. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2510.09717. arXiv preprint arXiv:2510.09717.

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

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719/.

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

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172, 2023.

Negin Raoof, Litu Rout, Giannis Daras, Sujay Sanghavi, Constantine Caramanis, Sanjay Shakkottai, and Alex Dimakis. Infilling score: A pretraining data detection algorithm for large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9QPH1YQCMn.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, 2017.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, pages 2615–2632, 2021.

Gaurav Srivastava, Aafiya Shamshad Hussain, Zhenyu Bi, Swastik Roy, Priya Pitre, Meng Lu, Morteza Ziyadi, and Xuan Wang. Beyondbench: Contamination-resistant evaluation of reasoning in language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mIKqVWGjwI.

John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002.

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

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

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Testing exchangeability on-line. In *International Conference on Machine Learning*, pages 768–775. AAAI Press, 2003.

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

Thuy-Trang Vu, Xuanli He, Gholamreza Haffari, and Ehsan Shareghi. Koala: An index for quantifying overlaps with pre-training corpora. In *EMNLP (Demos)*, pages 90–98. Association for Computational Linguistics, 2023.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. ReCaLL: Membership inference via relative conditional log-likelihoods. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.493. URL https://aclanthology.org/2024.emnlp-main.493/.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022.

Jiatong Yi and Yanyang Li. Membership inference on llms in the wild, 2026. URL https://arxiv.org/abs/2601.11314.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. Detecting data contamination in LLMs via in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YlpaaYxx4t.

Hengxiang Zhang, Songxin Zhang, Bingyi Jing, and Hongxin Wei. Fine-tuning can help detect pretraining data from large language models. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=X8dzvdkQw0.

Huixuan Zhang, Yun Lin, and Xiaojun Wan. PaCoST: Paired confidence significance testing for benchmark contamination detection in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1794–1809, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.97. URL https://aclanthology.org/2024.findings-emnlp.97/.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=ZGkfoufDaU.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining data detection for large language models: A divergence-based calibration method. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.300. URL https://aclanthology.org/2024.emnlp-main.300/.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, et al. Mmlu-cf: A contamination-free multi-task language understanding benchmark. *arXiv preprint arXiv:2412.15194*, 2024.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=gjfOL9z5Xr.

Wenhong Zhu, Hongkun Hao, Zhiwei He, Yun-Ze Song, Jiao Yueyang, Yumeng Zhang, Hanxu Hu, Yiran Wei, Rui Wang, and Hongyuan Lu. CLEAN–EVAL: Clean evaluation on contaminated large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 835–847, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.53. URL https://aclanthology.org/2024.findings-naacl.53/.

A Table of Notations

Table 3: Summary of notations used in the paper.

Notation	Description
θ_k	The k-th audited language model, $k \in \{1, \dots, K\}$.
K	The number of audited models.
$\mathcal{D}_{\mathrm{train}}$	The private training corpus, written $\mathcal{D}_{\mathrm{train}}(\theta_k)$ for audited model θ_k.
x_i	The i-th candidate benchmark item; the pool is $\{x_i\}_{i=1}^{n}$.
$\mathcal{D}_{\mathrm{cal}}$	The shared calibration set $\{x_\ell\}_{\ell=n+1}^{n+m}$ whose items are known members of every audited model.
n, m	The numbers of candidate and calibration items.
$M(x;\theta_k),\ M_i^k$	The binary membership indicator of x (resp. candidate i) with respect to θ_k, with $M_i^k = M(x_i;\theta_k)$.
$H_{0,i}, H_{1,i}$	The joint null and alternative hypotheses: contaminated for at least one audited model, and jointly pure for all audited models.
$T(x;\theta_k)$	The detection score for item x under model θ_k; smaller values indicate stronger evidence of non-membership.
p_i^k	The model-wise conformal p-value for candidate item i and model θ_k.
p_i^*	The aggregated max-p statistic used to test the joint null, $p_i^* = \max_k p_i^k$.
$\widetilde{p}_i$	The envelope-transformed value $\widehat{F}_{\mathrm{fit}}(p_i^*)$.
$\mathcal{S}$	The selected shared benchmark.
$\mathcal{S}_k$	The model-wise selected set for audited model θ_k.
$\mathrm{GCP}(\mathcal{S})$	The global contamination proportion (GCP), i.e., the realized fraction of contaminated items in $\mathcal{S}$.
$\mathrm{GCR}(\mathcal{S})$	The global contamination rate, i.e., $\mathbb{E}[\mathrm{GCP}(\mathcal{S})]$.
α	The target GCR level.
$\mathrm{Power}(\mathcal{S})$	The expected fraction of jointly pure candidate items selected into $\mathcal{S}$.
F_0, F_1	The CDFs of p_i^* conditional on $H_{0,i}$ (contaminated) and $H_{1,i}$ (jointly pure), respectively.
F_{mix}	The marginal CDF of p_i^* over the candidate pool, $F_{\mathrm{mix}} = \pi_0 F_0 + (1-\pi_0)F_1$.
$f_0, f_1, f_{\mathrm{mix}}$	The densities corresponding to $F_0, F_1, F_{\mathrm{mix}}$.
λ	The right-tail threshold used to fit the envelope for JECS.
$G_0, G_1, G_{\mathrm{mix}}$	The right-tail conditional CDFs on $[\lambda, 1]$ corresponding to $F_0, F_1, F_{\mathrm{mix}}$.
$g_0, g_1, g_{\mathrm{mix}}$	The corresponding right-tail conditional densities.
$\widehat{G}_n$	The empirical estimate of G_{mix} computed from $\{p_i^* : p_i^* > \lambda\}$.
$\widehat{g}(\lambda^+)$	The estimated right-boundary density at λ.
$\widehat{F}_{\mathrm{fit}}$	The fitted conservative envelope used to transform max-p values.
π_0	The population proportion of contaminated candidate items.
$\widehat{\pi}_0$	The estimated proportion of contaminated candidate items.
$\mathcal{P}, \mathcal{P}_0, \mathcal{P}_{\mathrm{cal}}, \mathcal{P}_{\mathrm{rem}}$	The benchmark candidate pool, clean block, calibration block, and remainder block used in the experimental split.
ρ	The per-model training fraction within the remainder block in the controlled experiments.

B Theoretical analysis

This appendix states the formal assumptions used and gives the proof. Let π_0 and $1-\pi_0$ denote the proportions of contaminated and jointly pure candidate items, respectively. Let $F_0, F_1, F_{\mathrm{mix}}$ and $f_0, f_1, f_{\mathrm{mix}}$ be the corresponding CDFs and densities of p_i^*.

B.1 Assumptions and their justification

Assumption 4 (Convex null max-p CDF)**.** The null CDF F_0 of p_i^* is convex on $[0,1]$.

Justification. The convexity condition is natural under max aggregation with heterogeneous ensemble exposure. For a contaminated item, let $\mathcal{K}_i^0 = \{k : M_i^k = 1\}, \quad c = |\mathcal{K}_i^0| \geq 1$ be the set and number of audited models whose training data contain item i. The remaining $K-c$ models treat the item as unseen and tend to produce small conformal p-values. Because JECS aggregates by $p_i^* = \max_k p_i^k$, these small values are masked in the right tail. Thus, on the operating interval $[\lambda, 1]$, the null CDF is governed by the maximum over the c models whose training data contain item i.

Two benchmark dependence regimes illustrate the shape of this CDF. If the null p-values from the c models whose training data contain item i are independent and marginally uniform, then

$$F_{0,c}(x) = \mathbb{P}\left(\max_{\ell \leq c} p_\ell \leq x\right) = x^c, \qquad x \in [\lambda, 1], \tag{19}$$

which is strictly convex for $c \geq 2$. At the other extreme, if these c null p-values are perfectly positively dependent, they move together and the maximum has CDF

$$F_{0,c}(x) = x, \tag{20}$$

which is linear and hence still convex. This case also satisfies the derivative condition used in the proof as an equality.

A positive-dependence model interpolates between these two extremes. Let $C_{\mathrm{co},c}$ denote the comonotonic copula, which represents perfect positive dependence, and let $C_{\mathrm{ind},c}$ denote the independence copula. Under perfect positive dependence, the c marginally uniform null p-values can be represented by a shared latent variable $Z \sim \mathrm{Unif}(0,1)$, namely $U_1 = \cdots = U_c = Z$. Hence

$$\mathbb{P}(U_1 \leq u_1, \ldots, U_c \leq u_c) = \mathbb{P}\left(Z \leq \min_{\ell \leq c} u_\ell\right) = \min_{\ell \leq c} u_\ell. \tag{21}$$

Therefore

$$C_{\mathrm{co},c}(u_1, \ldots, u_c) = \min_{\ell \leq c} u_\ell, \qquad C_{\mathrm{ind},c}(u_1, \ldots, u_c) = \prod_{\ell=1}^{c} u_\ell. \tag{22}$$

For $\eta \in [0,1]$, consider the convex combination

$$C_{\eta,c} = \eta C_{\mathrm{co},c} + (1-\eta) C_{\mathrm{ind},c}. \tag{23}$$

Evaluating this copula on the diagonal gives

$$F_{0,c}(x) = C_{\eta,c}(x, \ldots, x) = \eta x + (1-\eta) x^c. \tag{24}$$

Therefore

$$F''_{0,c}(x) = (1-\eta) c(c-1) x^{c-2} \geq 0 \tag{25}$$

for $c \geq 2$, while the cases $c = 1$ or $\eta = 1$ reduce to the linear CDF $F_{0,c}(x) = x$. If the exposure frequency varies across contaminated items, the overall null CDF is a mixture

$$F_0(x) = \sum_{c=1}^{K} w_c F_{0,c}(x), \qquad w_c \geq 0, \quad \sum_{c=1}^{K} w_c = 1, \tag{26}$$

and convexity is preserved under this mixture. This provides a structural explanation for Assumption 4: the max operator removes the influence of models whose training data do not contain the item in the right tail, and the dependence among models whose training data contain the item ranges from independence, which gives x^c, to perfect positive dependence, which gives x.

The mixture $C_{\eta,c}$ belongs to the positive Fréchet family of copulas [Genest et al., 2003, Koike et al., 2024], which models positive dependence as a convex combination of independence and comonotonicity. This family covers common relationships among audited models. For instance, knowledge distillation strengthens teacher–student alignment and moves $F_{0,c}$ toward the comonotonic endpoint x within this family, which preserves convexity. By contrast, sample-dependent interventions applied to only a subset of the audited models, such as targeted unlearning, can induce heterogeneous cross-model dependence outside this family, in which case the null max-p CDF need not be convex and Assumption 4 may fail.

Assumption 5 (Distinguishability of pure and contaminated items)**.** On the truncated interval $[\lambda, 1]$, the right-tail conditional CDF of pure items is at least that of contaminated items:

$$G_1(x) \geq G_0(x), \qquad x \in [\lambda, 1]. \tag{27}$$

Justification. The inequality is a distinguishability statement. Recalling that G_0, G_1 are right-tail conditional CDFs (Equation (13)), the assumption $G_1(x) \geq G_0(x)$ on $[\lambda, 1]$ is equivalent to

$$\Pr\big(p_i^* > x \,\big|\, p_i^* > \lambda,\ H_{1,i} \text{ true}\big) \ \leq\ \Pr\big(p_i^* > x \,\big|\, p_i^* > \lambda,\ H_{0,i} \text{ true}\big), \qquad x \in [\lambda, 1],$$

i.e., conditional on having crossed λ, a pure item is less likely than a contaminated item to escape further into the upper tail at every level x. Equivalently, the conditional law of p_i^* given $p_i^* > \lambda$ is stochastically smaller for pure items than for contaminated items, so the two populations remain distinguishable even after both have entered the right tail.

This assumption formalizes a one-sided separation within the right tail. Conditional on having crossed λ, pure items are expected to produce shallower exceedances, whereas contaminated items are expected to have a heavier upper tail. In the joint setting, this behavior is natural because an item contaminated in even a single audited model $\theta_{k^\star}$ satisfies $p_i^* = \max_k p_i^k \geq p_i^{k^\star}$ and cannot be pulled downward by small p-values from the remaining non-member models, whereas a jointly pure item must appear non-member-like across all models. Thus, among items with $p_i^* > \lambda$, pure items concentrate closer to λ, while contaminated items are more likely to extend toward 1.

More formally, a standard condition in hypothesis testing for characterizing separation between the null and alternative distributions is the monotone likelihood ratio property, which here requires the density ratio $f_1(x)/f_0(x)$ to be non-increasing so that larger max-p values provide stronger evidence of contamination than of joint purity. The following proposition shows that this condition is sufficient for Assumption 5.

Proposition 6 (Sufficient condition for Assumption 5)**.** *Suppose that* $1-F_0(\lambda) > 0$ *and* $1-F_1(\lambda) > 0$*, and that* $f_0(x) > 0$ *on* $[\lambda, 1]$ *with the likelihood ratio*

$$r(x) = \frac{f_1(x)}{f_0(x)} \tag{28}$$

non-increasing on $[\lambda, 1]$*. Then* $G_1(x) \geq G_0(x)$ *for all* $x \in [\lambda, 1]$*.*

Proof. For $j \in \{0, 1\}$, the right-tail conditional density on $[\lambda, 1]$ is $g_j(t) = f_j(t)/(1 - F_j(\lambda))$. Hence the conditional likelihood ratio

$$\tilde{r}(t) = \frac{g_1(t)}{g_0(t)} = \frac{1 - F_0(\lambda)}{1 - F_1(\lambda)}\, r(t) \tag{29}$$

is also non-increasing on $[\lambda, 1]$. Fix any $x \in [\lambda, 1]$. For all $u \in [\lambda, x]$ and $v \in [x, 1]$, monotonicity gives $\tilde{r}(u) \geq \tilde{r}(x) \geq \tilde{r}(v)$, and therefore

$$g_1(u)\, g_0(v) \geq g_0(u)\, \tilde{r}(x)\, g_0(v) \geq g_0(u)\, g_1(v). \tag{30}$$

Integrating this inequality over $u \in [\lambda, x]$ and $v \in [x, 1]$ yields

$$G_1(x)\big(1 - G_0(x)\big) \geq G_0(x)\big(1 - G_1(x)\big), \tag{31}$$

which simplifies to $G_1(x) \geq G_0(x)$. □

Proposition 7. *Suppose that the candidate maxima* p_i^* *are i.i.d. from* F_{mix}*. The threshold* $\lambda \in (0, 1)$ *is fixed,* $1 - F_{\mathrm{mix}}(\lambda) > 0$*, and the conditional right-tail distribution*

$$G_{\mathrm{mix}}(x) = \frac{F_{\mathrm{mix}}(x) - F_{\mathrm{mix}}(\lambda)}{1 - F_{\mathrm{mix}}(\lambda)}, \qquad x \in [\lambda, 1], \tag{32}$$

has a positive right-boundary density $g_{\mathrm{mix}}(\lambda^+)$ that is continuous in a right neighborhood of λ. Let $m_R = \sum_{i=1}^n \mathbf{1}\{p_i^* > \lambda\}$. The kNN neighborhood size k_n satisfies

$$k_n \to \infty, \qquad \frac{k_n}{m_R} \xrightarrow{p} 0. \tag{33}$$

Let $Y_{(1)} \le \cdots \le Y_{(m_R)}$ be the order statistics of the right-tail sample $\{p_i^* : p_i^* > \lambda\}$. Define

$$\widehat{g}(\lambda^+) = \frac{k_n}{m_R\,(Y_{(k_n)} - \lambda)}, \qquad \widehat{G}_n(x) = \frac{1}{m_R}\sum_{i=1}^n \mathbf{1}\{\lambda < p_i^* \le x\}, \quad x \in [\lambda, 1]. \tag{34}$$

Then

$$\widehat{g}(\lambda^+) \xrightarrow{p} g_{\mathrm{mix}}(\lambda^+), \qquad \sup_{x\in[\lambda,1]} |\widehat{G}_n(x) - G_{\mathrm{mix}}(x)| \xrightarrow{p} 0. \tag{35}$$

Proof. Since $1 - F_{\mathrm{mix}}(\lambda) > 0$, the law of large numbers gives $m_R/n \xrightarrow{p} 1 - F_{\mathrm{mix}}(\lambda)$, hence $m_R \xrightarrow{p} \infty$. Conditional on m_R, the right-tail sample is i.i.d. from G_{mix}, so the Glivenko–Cantelli theorem gives the uniform convergence of $\widehat{G}_n$ on $[\lambda, 1]$.

It remains to check the boundary density estimator. Let $U_n = G_{\mathrm{mix}}(Y_{(k_n)})$. By the standard uniform order-statistic representation, $U_n/(k_n/m_R) \xrightarrow{p} 1$ when $k_n \to \infty$ and $k_n/m_R \to 0$. Continuity and positivity of $g_{\mathrm{mix}}(\lambda^+)$ imply

$$G_{\mathrm{mix}}(\lambda + r) = g_{\mathrm{mix}}(\lambda^+)r + o(r), \qquad r \downarrow 0. \tag{36}$$

With $r_n = Y_{(k_n)} - \lambda$, the order-statistic result gives $G_{\mathrm{mix}}(\lambda + r_n) = U_n \to 0$ in probability, hence $r_n \to 0$ in probability and

$$\widehat{g}(\lambda^+) = \frac{k_n/m_R}{r_n} = \frac{k_n/m_R}{G_{\mathrm{mix}}(\lambda + r_n)}\frac{G_{\mathrm{mix}}(\lambda + r_n)}{r_n} \xrightarrow{p} g_{\mathrm{mix}}(\lambda^+). \tag{37}$$

□

Lemma 8 (Right-tail boundary dominance)**.** *Under Assumption 5, the mixture right-tail boundary density satisfies $g_{\mathrm{mix}}(\lambda^+) \ge g_0(\lambda^+)$.*

Proof. Write the right-tail mixture density as $g_{\mathrm{mix}}(\lambda^+) = w\, g_0(\lambda^+) + (1 - w)g_1(\lambda^+)$, where $w = \pi_0(1 - F_0(\lambda))/(1 - F_{\mathrm{mix}}(\lambda)) \in [0, 1]$. Assumption 5 implies $g_1(\lambda^+) \ge g_0(\lambda^+)$ at the boundary, so the convex combination is at least $g_0(\lambda^+)$. □

Lemma 9 (Asymptotic GCR control of Storey-BH; Storey et al., 2004, Thm. 4)**.** *Consider n tests $\{H_{0,i}\}_{i=1}^n$ with input p-values $\{p_i\}_{i=1}^n \subset [0, 1]$ and order statistics $p_{(1)} \le \cdots \le p_{(n)}$. Recall that p_i is* valid *under $H_{0,i}$ if $\mathbb{P}(p_i \le t \mid H_{0,i}) \le t$ for all $t \in [0, 1]$, as in Equation (9); assume the inputs are valid in this sense. Let $n_0 = |\{i : H_{0,i}$ is true$\}|$ denote the number of true nulls, and let $\pi_0 \in (0, 1]$ denote the limiting null proportion, $n_0/n \to \pi_0$. Let $\widehat{F}_0$ and $\widehat{F}$ denote the empirical null and overall CDFs of $\{p_i\}_{i=1}^n$, and assume both converge in probability, uniformly in $t \in [0, 1]$, to continuous limits. Let $\widehat{\pi}_0$ satisfy $\widehat{\pi}_0 \xrightarrow{p} \pi_0^\infty$ for some constant $\pi_0^\infty \ge \pi_0$. For $\alpha \in (0, 1)$, the Storey-BH selection*

$$\mathcal{S} = \left\{ i : p_i \le \frac{\alpha\, r^*}{\widehat{\pi}_0\, n} \right\}, \text{ where } r^* = \max\left\{ r : p_{(r)} \le \frac{\alpha\, r}{\widehat{\pi}_0\, n} \right\}, \tag{38}$$

with $r^* = 0$ and $\mathcal{S} = \emptyset$ when the set is empty, satisfies

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{1}\{i \in \mathcal{S},\ H_{0,i} \text{ is true}\}}{1 \vee |\mathcal{S}|} \right] \le \alpha. \tag{39}$$

B.2 Proof of Proposition 2

Proof. The conclusion uses Assumption 4 and Assumption 5. We first establish the bound at $x = \lambda$. The map $h(y) = \lambda y/(1 + \lambda y)$ is increasing, so Lemma 8 gives

$$F_{\text{fit}}(\lambda) \geq \frac{\lambda g_0(\lambda^+)}{1 + \lambda g_0(\lambda^+)} = \frac{\lambda f_0(\lambda^+)}{1 - F_0(\lambda) + \lambda f_0(\lambda^+)}. \tag{40}$$

Let $A = F_0(\lambda) \in [0, 1)$ and $B = \lambda f_0(\lambda^+)$. By the convexity of F_0 and $F_0(0) = 0$, the secant slope $F_0(\lambda)/\lambda$ is at most the right derivative $f_0(\lambda^+)$, so $B \geq A$. Hence $B(1 - A) \geq A(1 - A)$ and $B/(1 - A + B) \geq A$; therefore $F_{\text{fit}}(\lambda) \geq F_0(\lambda)$.

For $x \in (0, \lambda]$, the convexity of F_0 on $[0, \lambda]$ with $F_0(0) = 0$ implies that $F_0(x)/x$ is nondecreasing, so $F_0(x) \leq F_0(\lambda)x/\lambda$. Combined with the bound at λ,

$$F_{\text{fit}}(x) = \frac{F_{\text{fit}}(\lambda)}{\lambda} x \geq \frac{F_0(\lambda)}{\lambda} x \geq F_0(x). \tag{41}$$

For $x \in (\lambda, 1]$, write $F_0(x) = F_0(\lambda) + (1 - F_0(\lambda))G_0(x)$. Assumption 5 gives $G_{\text{mix}}(x) \geq G_0(x)$, and we just showed $F_{\text{fit}}(\lambda) \geq F_0(\lambda)$. The map $(a, b) \mapsto a + (1 - a)b$ is nondecreasing in $a \in [0, 1)$ for fixed $b \in [0, 1]$ and nondecreasing in b for fixed a, so

$$F_{\text{fit}}(x) = F_{\text{fit}}(\lambda) + \big(1 - F_{\text{fit}}(\lambda)\big)G_{\text{mix}}(x) \geq F_0(\lambda) + \big(1 - F_0(\lambda)\big)G_0(x) = F_0(x). \tag{42}$$

□

B.3 Proof of Theorem 3

By Proposition 7, the fitted envelope converges to F_{fit} uniformly on the two pieces of Equation (15). By Proposition 2, F_{fit} dominates F_0. For a null item and any $t \in (0, 1)$,

$$\mathbb{P}(F_{\text{fit}}(p_i^*) \leq t \mid H_{0,i}) = \mathbb{P}(p_i^* \leq F_{\text{fit}}^{-1}(t) \mid H_{0,i}) \leq \mathbb{P}(p_i^* \leq F_0^{-1}(t) \mid H_{0,i}) \leq t, \tag{43}$$

so the transformed null values are asymptotically super-uniform. The estimator Equation (16) is asymptotically conservative because

$$\widehat{\pi}_0 \xrightarrow{p} \frac{1 - F_{\text{mix}}(\lambda)}{1 - F_{\text{fit}}(\lambda)} \geq \frac{\pi_0(1 - F_0(\lambda))}{1 - F_0(\lambda)} = \pi_0, \tag{44}$$

using $F_{\text{fit}}(\lambda) \geq F_0(\lambda)$.

We now invoke Lemma 9 on the transformed inputs $\widetilde{p}_i = \widehat{F}_{\text{fit}}(p_i^*)$ with the selection in Equation (17). Equation (43) verifies the validity of $\{\widetilde{p}_i\}$ under the null (in the asymptotic sense, since $\widehat{F}_{\text{fit}} \to F_{\text{fit}}$ uniformly by Proposition 7). The Glivenko–Cantelli theorem applied to the i.i.d. null sub-sample and to the full i.i.d. sample $\{p_i^*\} \sim F_{\text{mix}}$, combined with the monotonicity of $\widehat{F}_{\text{fit}}$, gives the required uniform convergence of the null and overall empirical CDFs. The conservative null-proportion limit is supplied by Equation (44). Lemma 9 therefore yields $\limsup_{n\to\infty} \text{GCR}(\mathcal{S}) \leq \alpha$. □

C Algorithm Summary for JMCS

Algorithm 2 summarizes JMCS, the max-p baseline used in our experiments. It computes model-wise conformal p-values, aggregates them by $p_i^* = \max_k p_i^k$, and applies BH to the aggregated values to return the shared selected benchmark.

D Experimental details

Our experiments obtain ground-truth membership labels by controlled fine-tuning. For each audited model, an item is labeled as a member if it is included in that model's fine-tuning corpus, and as a non-member otherwise.

Algorithm 2 Joint Max Conformal Selection (JMCS)

Require: Candidate items $\{x_i\}_{i=1}^n$, audited models $\{\theta_k\}_{k=1}^K$, calibration data $\mathcal{D}_{\text{cal}} = \{x_i\}_{i=n+1}^{n+m}$, target level α.

1: **for** $k = 1, \dots, K$ **do**
2: Construct conformal p-values $\{p_i^k\}_{i=1}^n$ via Equation (4).
3: **end for**
4: Aggregate $p_i^* = \max_{1 \le k \le K} p_i^k$ for $i = 1, \dots, n$.
5: Sort the maxima $p_{(1)}^* \le \cdots \le p_{(n)}^*$ and compute $r^* = \max\{r : p_{(r)}^* \le \alpha\, r/n\}$.
6: Return $\mathcal{S} = \{i : p_i^* \le \alpha\, r^*/n\}$.

D.1 Tail-threshold selection

In all default runs of JECS, we choose the right-tail threshold λ from the grid $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. The criterion is to minimize the fitted left-branch slope. For each candidate λ, we estimate $\widehat{g}(\lambda^+)$ from the right-tail sample $\{p_i^* : p_i^* > \lambda\}$ and form the anchor

$$\widehat{F}_{\text{fit}}(\lambda) = \frac{\lambda\, \widehat{g}(\lambda^+)}{1 + \lambda\, \widehat{g}(\lambda^+)}. \tag{45}$$

We then set

$$\widehat{\lambda} = \arg\min_{\lambda} \frac{\widehat{F}_{\text{fit}}(\lambda)}{\lambda} = \arg\min_{\lambda} \frac{\widehat{g}(\lambda^+)}{1 + \lambda\, \widehat{g}(\lambda^+)}. \tag{46}$$

This rule selects the threshold whose fitted left branch has the smallest slope. The selected $\widehat{\lambda}$ is then used in Equation (15) to construct $\widehat{F}_{\text{fit}}$, transform the max-p values, and estimate $\widehat{\pi}_0$.

D.2 Dataset split setting

This subsection explains how we construct the controlled member/non-member structure that supplies the ground-truth joint-contamination labels used in our experiments.

Starting from the candidate pool $\mathcal{P}$ of a benchmark, we split the items into three disjoint parts: a *clean block* $\mathcal{P}_0$, a *calibration block* $\mathcal{P}_{\text{cal}}$, and a *remainder block* $\mathcal{P}_{\text{rem}}$, with sizes proportional to a, b, and $1 - a - b$ respectively. For each item x_i and each audited model θ_k, the membership label $M_i^k \in \{0, 1\}$ is then drawn block-by-block:

$$\mathbb{P}(M_i^k = 0) \;=\; \begin{cases} 1, & x_i \in \mathcal{P}_0, \\ 0, & x_i \in \mathcal{P}_{\text{cal}}, \\ 1 - \rho, & x_i \in \mathcal{P}_{\text{rem}}. \end{cases} \tag{47}$$

The three blocks play distinct roles. Items in $\mathcal{P}_0$ are never used to fine-tune any audited model, so every item in this block is jointly pure by construction. Items in $\mathcal{P}_{\text{cal}}$ are included in the fine-tuning corpus of *every* audited model and are reused as the shared calibration set $\mathcal{D}_{\text{cal}}$ in Section 3.2. Items in $\mathcal{P}_{\text{rem}}$ supply the per-model contamination. For each audited model θ_k, we independently sample a fraction ρ of the remainder uniformly at random and add those items to θ_k's fine-tuning corpus. We call ρ the *training fraction*; it is the per-model sampling fraction *within the remainder block*, not within the full pool. Because the assignments on $\mathcal{P}_{\text{rem}}$ are independent across the K audited models, the joint-purity probability $\mathbb{P}(\bigvee_{k=1}^K M_i^k = 0)$ equals 1 on $\mathcal{P}_0$ and $(1 - \rho)^K$ on $\mathcal{P}_{\text{rem}}$. Selection is performed on $\mathcal{D}_{\text{test}} := \mathcal{P}_0 \cup \mathcal{P}_{\text{rem}}$, with the calibration block held out. Throughout the main evaluation we use $a = b = 0.3$ and $\rho = 0.125$, and we vary ρ in the sensitivity analysis of Section 4.3.

Evaluation protocol. All reported GCP and Power values are Monte-Carlo means over 500 repetitions on the fixed three-block partition of Appendix D.2. In each repetition we draw 80% of $\mathcal{D}_{\text{cal}}$ and 80% of $\mathcal{D}_{\text{test}}$ uniformly at random; the three-block split itself is held fixed across repetitions.

D.3 Fine-tuning setting

We fine-tune all audited models using LoRA-SFT. The LoRA adapter is configured with rank $r = 16$, scaling parameter $\alpha_{\text{LoRA}} = 32$, dropout 0.05, no bias terms, and task type `CAUSAL_LM`. We train all

Table 4: Dataset split statistics at $\rho = 0.125$. Counts are reported for a single audited model θ_k: "Member" denotes $M_i^k = 1$ and "Non-member" denotes $M_i^k = 0$. Calibration items are shared members of all audited models, and the per-model test-set counts are identical across k.

Dataset	Component	Member	Non-member	Total
WikiMIA	Test set $\mathcal{D}_{\text{test}}$	39	514	553
	Calibration set $\mathcal{D}_{\text{cal}}$	236	0	236
	Candidate pool $\mathcal{P}$	275	514	789
ArXivTection	Test set $\mathcal{D}_{\text{test}}$	39	512	551
	Calibration set $\mathcal{D}_{\text{cal}}$	235	0	235
	Candidate pool $\mathcal{P}$	274	512	786
HackerNews	Test set $\mathcal{D}_{\text{test}}$	50	650	700
	Calibration set $\mathcal{D}_{\text{cal}}$	300	0	300
	Candidate pool $\mathcal{P}$	350	650	1000
DM-Math	Test set $\mathcal{D}_{\text{test}}$	50	650	700
	Calibration set $\mathcal{D}_{\text{cal}}$	300	0	300
	Candidate pool $\mathcal{P}$	350	650	1000
GitHub	Test set $\mathcal{D}_{\text{test}}$	50	650	700
	Calibration set $\mathcal{D}_{\text{cal}}$	300	0	300
	Candidate pool $\mathcal{P}$	350	650	1000
Pile-CC	Test set $\mathcal{D}_{\text{test}}$	50	650	700
	Calibration set $\mathcal{D}_{\text{cal}}$	300	0	300
	Candidate pool $\mathcal{P}$	350	650	1000
PubMed Central	Test set $\mathcal{D}_{\text{test}}$	50	650	700
	Calibration set $\mathcal{D}_{\text{cal}}$	300	0	300
	Candidate pool $\mathcal{P}$	350	650	1000

audited models with AdamW using learning rate 10^{-3}, weight decay 5×10^{-4}, 10 warmup steps, a cosine learning-rate schedule, and per-device batch size 8 for both training and evaluation. The number of fine-tuning epochs is backbone-specific: GPT-NeoX-20B is trained for 5 epochs, while the Pythia and LLaMA backbones are trained for 3 epochs.

D.4 Computational cost

All audited models are fine-tuned with the LoRA-SFT protocol of Appendix D.3, using PyTorch FSDP on NVIDIA RTX 4090 GPUs. Fine-tuning one audited model on one dataset takes about 5 GPU-min for LLaMA-7B or Pythia-6.9B (3 epochs, 2 GPUs), and about 15 GPU-min for GPT-NeoX-20B (5 epochs, 4 GPUs). After the scores $T(x; \theta_k)$ are extracted, JECS only performs NumPy operations on the (n_{test}, K) score matrix, including conformal p-value computation, max-p aggregation, envelope fitting, Storey estimation, and Storey-BH; every evaluated setting in Section 4 runs in under one minute on a single CPU core and requires no GPU.

E Additional Experimental Results

E.1 Synthetic setup for naive composition failure

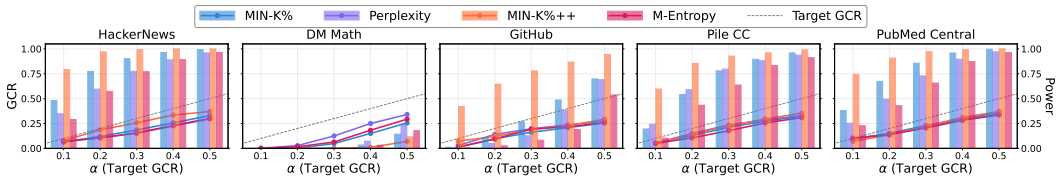
We generate a pool of 1200 candidate items for $K = 4$ audited models $\{\theta_k\}_{k=1}^K$, and split it into a calibration set $\mathcal{D}_{\text{cal}}$ of size $n_{\text{cal}} = 360$ (30% of the pool) and an audit set $\mathcal{D}_{\text{test}}$ of size $n_{\text{test}} = 840$. Items in $\mathcal{D}_{\text{cal}}$ are members of every audited model by construction, matching the shared-member calibration assumption of Section 3.2. Items in $\mathcal{D}_{\text{test}}$ receive independent Bernoulli labels $M_i^k \overset{\text{iid}}{\sim} \text{Bernoulli}(0.30)$ across both i and k, so the joint-purity probability of an audit item is $(1 - 0.30)^K \approx 0.24$. For each item i and audited model θ_k, the detection score is

$$T_i^k = \mu \cdot M_i^k + \varepsilon_i^k, \qquad \varepsilon_i^k \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \tag{48}$$

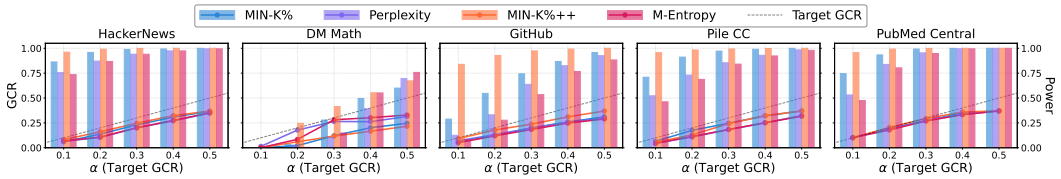
with signal strength $\mu = 4$, so members and non-members are $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(0, 1)$, and the model-wise scores are independent across k.

E.2 MIMIR Subsets

We further evaluate JECS on five subsets (HackerNews, DMMath, GitHub, Pile-CC, PubMed) of the MIMIR benchmark [Duan et al., 2024], fine-tuning $K = 8$ audited models from Pythia-6.9B and LLaMA-7B per subset under the same protocol as Section 4.1. Figure 9 reports realized GCP and Power for the four detection scores in Section 4.1 at $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The realized GCP stays at or below the target diagonal throughout, confirming that GCR control extends beyond the WikiMIA and ArXivTection candidate pools used in Section 4.



(a) Pythia-6.9B



(b) LLaMA-7B

Figure 9: Realized GCP and Power of JECS on five MIMIR subsets at $K = 8$ across Pythia-6.9B and LLaMA-7B. GCP curves (lines, left axis) and Power bars (right axis) for four detection scores at $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ on HackerNews, DMMath, GitHub, Pile-CC, and PubMed; the dashed line is the target GCR.

E.3 Power on WikiMIA

Table 5 reports Power for JECS and JMCS on WikiMIA at $K = 16$, mirroring the ArXivTection table of Table 2. JECS attains higher Power than JMCS across the evaluated targets, model families, and detection scores, with the largest gain on LLaMA-7B with Perplexity at $\alpha = 0.3$, where Power rises from 0.088 to 0.439. This pattern matches the aggregate finding of Section 4 and the super-uniformity tax interpretation of Section 3.2.

E.4 Cross-experiment variability of JECS on Min-K%++

Figure 10 shows the cross-experiment variability of JECS with Min-K%++ at $K = 16$. The GCP remains aligned with the target diagonal, and the shaded band is narrow, with its upper edge only slightly above the target at smaller α. Power also varies little, supporting the stability of the gains reported in Table 2.

E.5 Mixed-family audit pool

The main-text protocol audits K fine-tuned models initialized from a single backbone, which leaves open whether JECS's control depends on the audited models sharing an architecture and parameter scale. We complement that protocol with a heterogeneous audit pool composed of $K = 16$ fine-tuned variants drawn from 8 public LLM configurations spanning three families and a wide scale range: Pythia [Biderman et al., 2023] at 1.4B, 2.8B, 6.9B, and 12B, GPT-NeoX-20B [Black et al., 2022], and LLaMA [Touvron et al., 2023] at 7B, 13B, and 30B. Each of the 8 configurations contributes two independently fine-tuned models over disjoint contamination assignments on the

Table 5: Power on WikiMIA at $K = 16$. Reported values are average Power with standard error. "Base" refers to JMCS and "Ours" refers to JECS. **Bold** numbers mark the better result in each pair.

Model	Score	$\alpha = 0.1$		$\alpha = 0.2$		$\alpha = 0.3$	
		Base	Ours	Base	Ours	Base	Ours
NeoX-20B	Perplexity	0.000±0.000	**0.094**±0.005	0.089±0.003	**0.394**±0.004	0.326±0.002	**0.565**±0.005
	Min-K%++	0.382±0.001	**0.443**±0.001	0.499±0.001	**0.574**±0.002	0.580±0.001	**0.643**±0.001
	Min-K%	0.006±0.001	**0.067**±0.003	0.152±0.003	**0.358**±0.003	0.358±0.001	**0.529**±0.003
	M-Entropy	0.000±0.000	**0.017**±0.002	0.041±0.002	**0.283**±0.004	0.229±0.003	**0.460**±0.005
LLaMA-7B	Perplexity	0.000±0.000	**0.010**±0.001	0.010±0.001	**0.171**±0.005	0.088±0.002	**0.439**±0.007
	Min-K%++	0.120±0.005	**0.400**±0.002	0.456±0.001	**0.527**±0.001	0.526±0.001	**0.592**±0.002
	Min-K%	0.018±0.001	**0.133**±0.002	0.178±0.001	**0.316**±0.002	0.301±0.001	**0.469**±0.003
	M-Entropy	0.000±0.000	**0.006**±0.001	0.007±0.001	**0.130**±0.004	0.064±0.002	**0.363**±0.006
Pythia-6.9B	Perplexity	0.000±0.000	**0.008**±0.001	0.012±0.001	**0.048**±0.002	0.043±0.001	**0.168**±0.006
	Min-K%++	0.000±0.000	**0.003**±0.001	0.010±0.001	**0.232**±0.007	0.181±0.004	**0.486**±0.003
	Min-K%	0.000±0.000	**0.001**±0.000	0.003±0.001	**0.036**±0.003	0.029±0.002	**0.311**±0.007
	M-Entropy	0.000±0.000	**0.004**±0.001	0.011±0.001	**0.034**±0.002	0.031±0.001	**0.125**±0.005

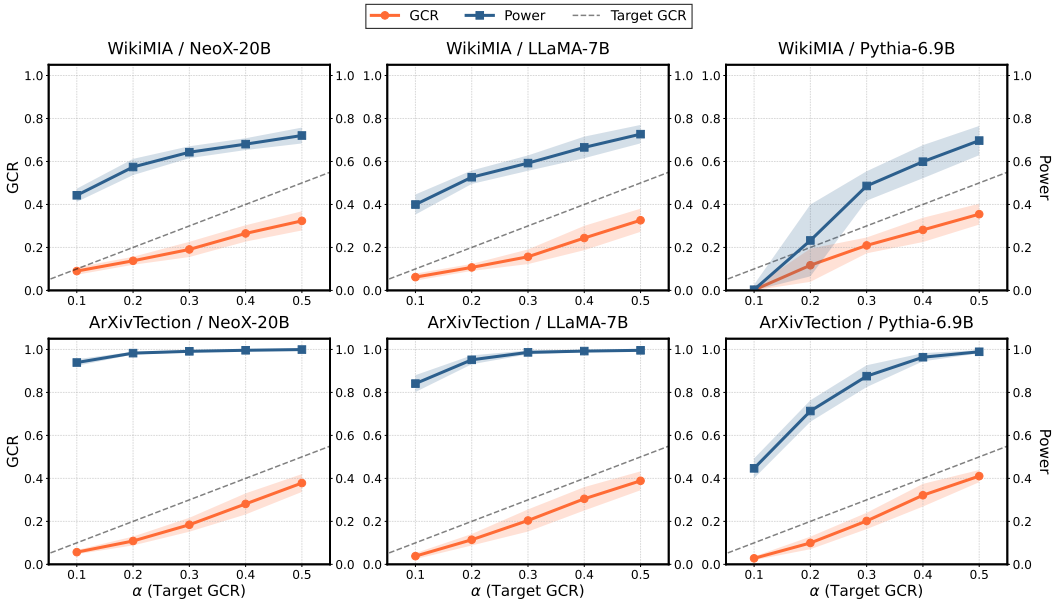


Figure 10: JECS with Min-K%++ at $K = 16$: cross-experiment dispersion. Curves report mean realized GCP (left axis, orange) and Power (right axis, blue) on WikiMIA (top row) and ArXivTection (bottom row), for GPT-NeoX-20B, LLaMA-7B, and Pythia-6.9B. Shaded bands are ± 1 standard deviation, clipped to $[0, 1]$; the dashed diagonal is the target GCR.

remainder block $\mathcal{P}_{\text{rem}}$, so the audit pool varies architecture, parameter scale, and per-model training data jointly. The shared-member calibration block $\mathcal{P}_{\text{cal}}$ from Appendix D.2 is reused by every variant unchanged; the rest of the protocol — test split, four detection scores, and the target grid $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ — matches Section 4.1.

Figure 11 reports realized GCP and Power on this mixed-family pool for the four detection scores at $K = 16$, on WikiMIA and ArXivTection. The realized GCP stays at or below the target diagonal throughout; Min-K%++ delivers the highest Power on ArXivTection while the four scores cluster more tightly on WikiMIA. The picture matches Section 4: JECS controls GCR without score-specific tuning and without architecture- or scale-matched audited models.

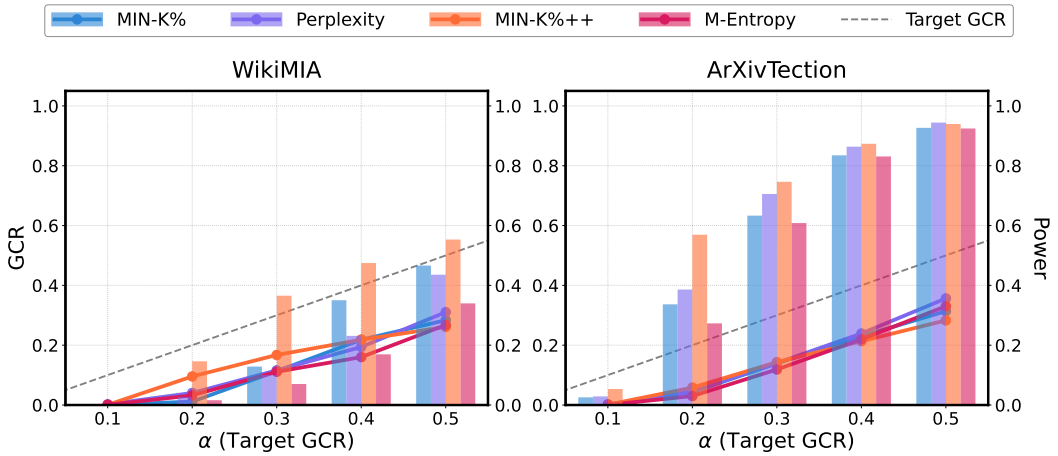


Figure 11: JECS on a mixed-family audit pool at $K = 16$. Realized GCP (lines, left axis) and Power (bars, right axis) for four detection scores at $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ on WikiMIA (left) and ArXivTection (right). The $K = 16$ pool combines 8 fine-tuned configurations spanning Pythia (1.4B/2.8B/6.9B/12B), GPT-NeoX-20B, and LLaMA (7B/13B/30B), with two independently fine-tuned models per configuration. The dashed line is the target GCR.

E.6 Recent open-weight model

To evaluate JECS on a recently released model, we audit $K = 16$ LoRA adapters fine-tuned from Qwen3.5-9B-Base, using the checkpoint revision published on March 2, 2026. We construct a post-release benchmark from a fixed snapshot of the Hugging Face arXiv metadata, selecting 800 deduplicated abstracts of 32 to 256 words whose first version was submitted between May 1 and June 30, 2026. These abstracts postdate the checkpoint, so the base model is unlikely to have seen them during pretraining and membership is governed by our controlled fine-tuning. Following the three-block split of Appendix D.2, we use 240 clean items, 240 shared calibration members, and a remainder of 320 items from which each adapter samples 40, which corresponds to $\rho = 0.125$. Each adapter uses the LoRA rank, scaling, and dropout of Appendix D.3 with learning rate 10^{-3} and batch size 8. Because Qwen3.5 interleaves full and linear attention layers, the LoRA targets also include the linear-attention projections. Detection scores are extracted from the epoch-3 checkpoint, and evaluation follows the protocol of Appendix D.2.

Table 6 shows that the empirical GCR of JECS stays below α for all four detection scores and target levels. Power exceeds 0.98 throughout because the fine-tuned members are strongly separable from post-release abstracts under this model.

Table 6: Empirical GCR and Power of JECS on post-release arXiv abstracts with $K = 16$ Qwen3.5-9B-Base adapters.

Score	$\alpha = 0.1$		$\alpha = 0.2$		$\alpha = 0.3$	
	GCR	Power	GCR	Power	GCR	Power
Perplexity	0.061	0.989	0.143	0.996	0.227	0.996
Min-K%++	0.054	1.000	0.160	1.000	0.241	1.000
Min-K%	0.058	0.995	0.140	0.997	0.229	0.997
M-Entropy	0.064	0.984	0.142	0.995	0.224	0.996

F Feasibility of Constructing the Shared-Member Calibration Set

The calibration assumption in Section 3.2 requires a pool of items that are shared members of every audited model. We argue this is achievable in practice on three independent grounds.

1. **Public benchmarks expose training splits that are routinely mixed into LLM training.** Public benchmarks release explicit training splits, and the contamination literature documents that such training data is in many cases deliberately incorporated into LLM pretraining or fine-tuning. For instance, OpenAI openly acknowledges mixing the MATH and GSM8K training sets into GPT-4's training [Achiam et al., 2023], and surveys catalog the same practice across model families including GLM-130B, Qwen, Nemotron-4, InternLM-2, and MiniCPM [Zhou et al., 2023, Xu et al., 2024]. Benchmark training splits thus constitute a sizable population of items that are members of many deployed models by construction.
2. **Model providers openly disclose canonical web sources as documented pretraining components.** Modern LLM families often identify canonical web sources as components of their pretraining mixes, including Common Crawl, Wikipedia, arXiv, and GitHub. Llama publishes the full breakdown (CommonCrawl, C4, GitHub, Wikipedia, arXiv, StackExchange) in its technical report [Touvron et al., 2023], and Falcon similarly releases the RefinedWeb extraction of Common Crawl [Penedo et al., 2023]. Decontamination audits further confirm substantial overlap between such corpora and widely used benchmarks [Yang et al., 2023], so samples drawn from these disclosed sources can be treated as shared members of the corresponding model families.
3. **Regulatory mandates are making training-data disclosure a legal requirement.** Beyond voluntary disclosure, Article 53(1)(d) of the EU AI Act (Regulation (EU) 2024/1689) requires providers of general-purpose AI (GPAI) models placed on the EU market, including OpenAI, Google, Anthropic, and others, to publish a "sufficiently detailed summary" of the content used for training, following a template from the EU AI Office [European Parliament and Council of the European Union, 2024]. These obligations apply to newly placed GPAI models from 2 August 2025. Enforcement powers, including fines of up to EUR 15 M or 3% of global annual turnover, start on 2 August 2026, and pre-existing GPAI models must comply by 2 August 2027. As these summaries become available, auditors will have more high-confidence evidence about which public sources and data collections were used across major deployed models, making the shared-member assumption increasingly easy to satisfy.

Each of the three grounds above provides realistic candidates for assembling the calibration pool in deployed audit settings. After collecting candidate shared members from these sources, the auditor restricts them to the benchmark domain and matches score-relevant properties such as topic, programming language, and sequence length before conformal calibration.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We summarize our contributions in both the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in Section 6

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The assumptions and proofs are provided in Appendix B

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We disclose all the information needed to reproduce the main experimental results.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Our code is available at https://github.com/ml-stat-Sustech/Joint-Pure-Benchmark-Selection.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We introduce the setup of our experiment in Appendix D.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: We report standard errors in Tables 2 and 5 and standard deviations in Figure 10.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The compute resources are provided in Appendix D.4

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics in every respect.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss the societal impacts of decontamination in the introduction.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: All models and datasets utilized in this study are publicly accessible.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All the creators or original owners of assets used in this study are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We release our code with documentation at `https://github.com/ml-stat-Sustech/Joint-Pure-Benchmark-Selection`.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

    Guidelines:

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