crBH: CONFORMAL TOP-k SELECTION WITH FALSE DISCOVERY RATE CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting candidates that truly belong to the top-k of a large pool is important in scientific screening and information retrieval, but their outcomes are unavailable at selection time. Existing conformal selection rules address fixed outcome thresholds; top-k membership instead depends on the other items in the test pool. Calibration and test ranks also live on different scales when their pool sizes differ. We formulate inductive top-k selection with false discovery rate (FDR) as its error criterion and propose crBH to address this rank-space mismatch. An exact rank projection supplies calibration-item probabilities of falling outside the test top-k; projected conformal statistics combine these probabilities with trained ranking scores. A null-proportion-adjusted BH step-up rule then selects a score-ordered subset. We characterize the asymptotic FDR and power of crBH and connect its power to the quality and ordering induced by the ranking model. Although the guarantee is asymptotic, controlled and application benchmarks show empirical FDR close to the nominal level at the studied finite sample sizes when the method makes substantial discoveries, while retaining substantial power. For applications requiring an exact finite-sample guarantee, we additionally give a conservative certified variant under joint calibration/test exchangeability.

1 INTRODUCTION

Selecting a subset of promising candidates from a large pool is crucial in various scientific and real-world applications, such as recruitment screening (Hunkenschroer & Luetge, 2022), recommendation systems (Yang et al., 2012), and information retrieval (Liu, 2009).

In these settings, the goal is often to identify candidates that truly belong to the top-k of a large pool. For example, in information retrieval, a search engine may retrieve a large set of candidate documents for a query, yet only a few results can be displayed to users, making it essential to reliably select those that are truly among the most relevant. While learning-based ranking algorithms have been extensively studied (Joachims, 2002; Burges et al., 2005; Burges, 2010), simply taking the model-predicted top-k candidates can be unreliable without formal guarantees. This raises the importance of developing rigorous top-k selection strategies with reliability guarantees.

Conformal selection provides a model-agnostic framework for selecting candidates whose unobserved outcomes exceed user-specified values. Its goal is to select as many truly qualifying units as possible while controlling the FDR, defined as the expected proportion of selected units that are false discoveries. Despite its distribution-free guarantee, conformal selection cannot be directly applied to the top-k selection problem by casting predicted rank as the unobserved outcome. This is because ranks are inherently sample-size dependent, so the outcome space for the calibration set can differ substantially from that of the test set when the two sets have different sizes. This raises a fundamental question: how to design a conformal selection method tailored to top-k selection?

In this work, we first study an inductive conformal top-k selection problem: Given a calibration set with observed internal calibration ranks and a trained ranking model, our goal is to select a subset of test items that maximizes the statistical power (i.e., the proportion of truly top-k items selected), while maintaining the FDR below a pre-specified level. To this end, we propose crBH, a novel and rigorous method that establishes a projection between the calibration and test rank spaces. Specifically, we first derive the probability distribution of each calibration item projected in the test ranking space via combinatorial constructions (Feller, 1968). Subsequently, we compute

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

projected statistics for test items based on the projected probability distributions of calibration items, indicating their item-wise statistical evidence for top-k membership. Finally, we select the subset using the BH procedure with a known proportion of non-top-k items. Our theory characterizes the asymptotic FDR and power of crBH and explains how ranking quality shapes discovery power.

The guarantee for crBH is asymptotic, while high-yield configurations show empirical FDR near the nominal level at the studied finite sample sizes. Controlled studies vary ranking quality, calibration size, target fraction, model class, nominal level, and rank projection to identify sources of power and finite-sample conservativeness. The counterexample in Proposition 7 rules out universal finite-sample validity for raw crBH; we also provide an optional conservative certificate under exchangeability.

We summarize our contributions as follows:

- We formulate inductive top-k selection with FDR as its error criterion and derive the exact projection between calibration and test rank spaces.
- We construct crBH from projected statistics and a null-proportion-adjusted BH step-up rule, characterize its asymptotic FDR and power, and identify how oracle ordering and ranking-model quality affect selection power.
- We evaluate crBH across controlled and application benchmarks; sensitivity, model-class, and projection studies explain where empirical power and conservativeness change.

As a theoretical safeguard, we also derive an optional conservative certification layer with exact finite-sample FDR control under joint exchangeability.

2 Related Work

Conformal selection. Conformal prediction quantifies uncertainty under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018; Angelopoulos & Bates, 2023), with extensions to adaptive prediction and risk control (Romano et al., 2019; 2020; Bates et al., 2021; Angelopoulos et al., 2025). Conformal selection uses such ideas for FDR-controlled decisions: Jin & Candès (2023) consider outcomes above fixed user-specified levels, while related work addresses outliers and other discovery tasks (Bates et al., 2023; Marandon, 2024). Their fixed-threshold nulls differ from our pool-dependent top-k hypotheses.

Learning to rank and rank uncertainty. Ranking methods serve retrieval, recommendation, and search (Manning et al., 2008; Baeza-Yates & Ribeiro-Neto, 2011; Ricci et al., 2015; Jannach et al., 2010; Brin & Page, 1998; Croft et al., 2009); representative models include RankSVM, RankNet, LambdaMART, and ListNet (Liu, 2009; Joachims, 2002; Burges et al., 2005; Burges, 2010; Cao et al., 2007). These models optimize ordering but do not control the error rate of a selected subset. Conformal ranking methods instead provide uncertainty sets or rank coverage (Fermanian et al., 2025; Liao et al., 2026). Our primary method crBH targets FDR-controlled inductive top-k selection with an asymptotic characterization; Section 4.4 gives an optional finite-sample certificate under exchangeability.

3 Problem Setup

We seek a reliable subset of test items belonging to the true top-k. Let $\mathcal{X}$ be the feature space and $\mathcal{Y} \subseteq \mathbb{R}$ the outcome space. Our theoretical results concern $n+m$ pairs $\{(X_i, Y_i)\}_{i=1}^{n+m}$ from a joint distribution $\mathcal{P}_{XY}$ under the following assumption.

Assumption 1 (Exchangeability)**.** The underlying sequence of random variables $\{(X_i, Y_i)\}_{i=1}^{n+m}$ is jointly exchangeable, conditional on any independent training information. Outcome ties are resolved by independent uniform labels, as specified below.

Assign each outcome an iid $V^Y \sim \mathrm{Unif}(0,1)$, independently of observations, training information and score tie labels, and use the lexicographic order of $\widetilde{Y} = (Y, V^Y)$. This preserves joint exchangeability and gives a strict order almost surely. Every outcome comparison below means this

augmented order; it agrees with the measured outcome order when there are no ties. With ties, the target is the resulting randomized exact test top-k set.

The outcome Y serves to induce the latent true ranking. Our inferential target is intrinsically rank-based: for each test item, we ask whether it belongs to the top-k set relative to the other candidates in the same test pool, rather than whether its outcome exceeds a fixed numerical threshold. Consequently, the target depends on the ordering induced by Y, but not on its absolute numerical scale.

In applications where calibration outcomes are observed, we convert them into internal calibration ranks and retain only this rank information for the conformal calibration step. In applications with ordinal supervision, these ranks may be provided directly. Thus, the required calibration interface is

$$R_i^c := 1 + \sum_{\ell=1}^{n} \mathbf{1}\{Y_\ell^c > Y_i^c\}, \qquad i \in [n], \tag{1}$$

where $[n] := \{1, \ldots, n\}$ and $[m] := \{1, \ldots, m\}$. We assume access to a calibration dataset $\mathcal{D}_{\mathrm{cal}} = \{(X_i^c, R_i^c)\}_{i=1}^n$. A ranking model may be trained on independent training data. The remaining m items form the test pool $\mathcal{D}_{\mathrm{test}}^X = \{X_j^t\}_{j=1}^m$, where both the true outcomes Y_j^t and their induced true test ranks $\{R_j^t\}_{j=1}^m$ are unobserved at selection time.

We consider a pre-trained scoring model $S : \mathcal{X} \to \mathbb{R}$ used to evaluate candidates. While the true ranking is determined by the unobserved Y, the *model-predicted ranking* is determined by sorting the predicted scores $S(X)$, where larger scores indicate a higher likelihood of belonging to the top tail. To ensure unambiguous score comparisons, we augment all scores with independent $\mathrm{Unif}(0, 1)$ tie-breakers, yielding lexicographically ordered scores $\bar{S}_i^c = (S(X_i^c), U_i^c)$ and $\bar{S}_j^t = (S(X_j^t), U_j^t)$.

Top-k selection. Fixing a target rank threshold k, our goal is to identify a subset of indices $\widehat{\mathcal{S}} \subseteq \{1, \ldots, m\}$ from the test pool that belong to the top-k, corresponding to testing the hypotheses

$$H_{0,j} : R_j^t > k \qquad \text{versus} \qquad H_{1,j} : R_j^t \le k, \qquad j \in [m]. \tag{2}$$

Let $\mathcal{H}_0 = \{j \in [m] : R_j^t > k\}$ denote the set of true null hypotheses. Since the test ranks form a permutation of $1, \ldots, m$, $|\mathcal{H}_0| = m - k$. Unlike evaluating outcomes against a static threshold, determining top-k membership is fundamentally a relative, pool-dependent rank event. Consequently, standard conformal inference tools cannot be directly applied to test these hypotheses.

We aim to identify the subset $\widehat{\mathcal{S}}$ such that the false discovery rate (FDR) is controlled at a user-specified target level $\alpha \in (0, 1)$:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{H}_0|}{\max(|\widehat{\mathcal{S}}|, 1)}\right] \le \alpha, \tag{3}$$

where the quantity inside the expectation is the false discovery proportion (FDP), representing the fraction of non-top-k items mistakenly included in the selected set. The FDR is the expected fraction of selected items that do not belong to the true test top-k set.

Remark 1 (Rank invariance)**.** For any strictly increasing function h, replacing every latent outcome Y by $h(Y)$ leaves the induced calibration and test ranks unchanged. Consequently, the top-k hypotheses, the FDR and power targets, and the inputs and outputs of the subsequent crBH procedure are invariant to strictly increasing transformations of the outcome scale.

Since an excessively low FDR can be trivially achieved by selecting very few items, it is also important to evaluate whether a method selects an adequate number of true top-k items. This aspect is quantified by the metric power:

$$\mathrm{Power} = \frac{1}{k}\mathbb{E}\left[|\widehat{\mathcal{S}} \setminus \mathcal{H}_0|\right]. \tag{4}$$

Subject to FDR control, power measures the ability of the procedure to recover true top-k items.

For $\pi = (m - k)/m$, selecting all test items has deterministic FDP π. Thus the branch $\alpha \ge \pi$ is immediately valid, including $k = m$. The nontrivial method and certification branches use $\alpha < \pi$.

4 METHOD

Fixed-threshold conformal selection rules (Bates et al., 2021; Jin & Candès, 2023; Marandon, 2024) are not directly applicable to pool-dependent top-k membership. We map calibration ranks into test rank space and use the projected null weights in crBH score comparisons.

4.1 THE crBH PROCEDURE

A fundamental challenge in our inductive setting is the size mismatch between the calibration ranking problem (size n) and the test ranking problem (size m). To motivate the construction, first suppose that the calibration and test pools have the same size, $n = m$. Then the indicator $N_i^c := \mathbf{1}\{R_i^c > k\}$ directly identifies whether calibration item i lies outside the calibration top-k set. Because the calibration and test ranks share the same scale, this indicator provides the natural calibration analogue of the test-pool null event for top-k selection.

When $n \neq m$, however, the observed calibration rank and the test top-k event live on different rank scales. Consequently, the indicator $\mathbf{1}\{R_i^c > k\}$ is not the appropriate null calibration label for the size-m test problem. We use two explicitly distinct auxiliary rank constructions. For $n \leq m$, place all n calibration ranks in a uniformly chosen size-n subset of m ordered slots, and let $\widetilde{R}_i$ be the slot of calibration rank R_i^c. For $n > m$, merge all n calibration labels with $m-1$ exchangeable competitor labels; $\widetilde{R}_i$ is the rank of item i among itself and those competitors. The first construction augments the entire calibration pool; the second compares one calibration item with new competitors. Their probability meanings are not interchangeable. Define the auxiliary projected null indicator

$$N_i^{\mathrm{proj}} := \mathbf{1}\{\widetilde{R}_i > k\}. \tag{5}$$

Because N_i^{proj} is not directly observed given R_i^c, we use its conditional expectation:

$$q_i^\star := \mathbb{E}\left[N_i^{\mathrm{proj}} \mid R_i^c\right] = \mathbb{P}\left(\widetilde{R}_i > k \mid R_i^c\right), \qquad i \in [n]. \tag{6}$$

The following theorem provides the exact closed-form expressions for $q_i^\star$.

Theorem 2 (Exact rank-projection probabilities)**.** *Let $r = R_i^c$ and fix $k \in \{1, \ldots, m\}$.*

$$q_i^\star = \sum_{x=\max(0,n-(m-k))}^{\min(r-1,k,n)} \frac{\binom{k}{x}\binom{m-k}{n-x}}{\binom{m}{n}}, \qquad n \leq m,$$

$$q_i^\star = \sum_{x=\max(0,r+k-m)}^{r-1} \frac{\binom{n}{x}\binom{m-1}{r+k-1-x}}{\binom{n+m-1}{r+k-1}}, \qquad n > m.$$

Writing $\pi = (m-k)/m$, these weights satisfy $\sum_{i=1}^n q_i^\star = n\pi$.

Appendix F derives both formulas; Appendix C.2 compares exact projection with a mass-matched scaled-rank approximation sharing its total mass and equal-size limit.

Thus, in (7), $q_i^\star$ supplies fractional null-reference weight: higher values contribute more null evidence. It is the conditional expectation of an auxiliary projected-null indicator, not a posterior probability that the calibration item is intrinsically null.

Remark 3 (Equal-size sanity check)**.** When $n = m$, projection does not change the rank scale, so $\widetilde{R}_i = R_i^c$ and $q_i^\star = \mathbf{1}\{R_i^c > k\}$. Thus, the projected null weights reduce to ordinary binary calibration null labels, and the subsequent projected statistic has the form of a binary-reference rank statistic. This reduction alone does not establish finite-sample validity for the pool-dependent nulls.

Definition 4 (projected conformal statistics)**.** For each test item $j \in \{1, \ldots, m\}$, define

$$p_j := \frac{1 + \sum_{i=1}^n q_i^\star \mathbf{1}\{\bar{S}_i^c > \bar{S}_j^t\}}{1 + \sum_{i=1}^n q_i^\star}. \tag{7}$$

The leading one is the conformal self-count. Here "conformal" denotes score comparison with calibration references, weighted for the pool-dependent top-k null; p_j is a projected statistic, not a universally finite-sample-valid p-value (Proposition 7).

Algorithm 1 crBH for top-k selection

Require: Calibration feature-rank data $\mathcal{D}_{\text{cal}} = \{(X_i^c, R_i^c)\}_{i=1}^n$, Test features $\mathcal{D}_{\text{test}}^X = \{X_j^t\}_{j=1}^m$, scoring model $S(\cdot)$, target FDR $\alpha \in (0,1)$, target rank $k \in \{1,\dots,m\}$.

1: **if** $k = m$ **then return** $\widehat{\mathcal{S}} = \{1,\dots,m\}$
2: **end if**
3: **Step 1: Compute exact rank-projection probabilities**
4: **for** $i = 1,\dots,n$ **do**
5: Read the observed internal calibration rank R_i^c.
6: Calculate $q_i^\star$ using the corresponding expression in Theorem 2.
7: **end for**
8: **Step 2: Compute projected conformal statistics**
9: **for** $j = 1,\dots,m$ **do**
10: Evaluate $p_j = \frac{1+\sum_{i=1}^n q_i^\star \mathbf{1}\{\bar{S}_i^c > \bar{S}_j^t\}}{1+\sum_{i=1}^n q_i^\star}$.
11: **end for**
12: **Step 3: Perform the null-proportion-adjusted BH step-up rule**
13: Calculate the adjusted target FDR: $\alpha' = \min\left(1, \frac{m}{m-k}\alpha\right)$.
14: Sort projected statistics in ascending order, breaking equal values by decreasing $\bar{S}$ order: $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$, and define $p_{(0)} := 0$.
15: Find the cutoff index: $\hat{\ell} = \max\left\{r \in \{0,\dots,m\} : p_{(r)} \le \frac{r}{m}\alpha'\right\}$.
16: **return** the first $\hat{\ell}$ test items in this order (empty if $\hat{\ell} = 0$).

Once the projected statistics are available, we turn to the step-up rule. In ordinary multiple testing, the null proportion $\pi_0 = |\mathcal{H}_0|/m$ is unknown. In the top-k problem, however, it is deterministic: exactly k test items are truly top-k, so exactly $m-k$ hypotheses are null and

$$\pi_0 = \frac{m-k}{m}.$$

This structural information allows us to use the π_0-corrected BH level

$$\alpha' = \min\left(1, \frac{\alpha}{\pi_0}\right) = \min\left(1, \frac{m}{m-k}\alpha\right),$$

leading to the crBH procedure in Algorithm 1.

The projected statistics are nondecreasing along the common decreasing score order, so Algorithm 1 returns a score prefix. A passing prefix cannot stop inside a block of equal statistics: the remaining equal values pass the larger critical value as well. The prefix formulation therefore agrees with the corresponding BH-style rejection set, while explicitly treating empty output. For $\alpha < \pi$ the unnormalized step-up condition is exactly

$$1 + \sum_i q_i^\star \mathbf{1}\{\bar{S}_i^c > s_{(r)}\} \le \alpha\left(n + \frac{1}{\pi}\right)\frac{r}{m}. \tag{8}$$

For $\alpha \ge \pi$, $\alpha' = 1$ and all projected statistics satisfy $p_j \le 1$, so all items are selected.

A visualization of the projected-statistic separation and selection cutoff appears in Appendix C.

4.2 Asymptotic characterization of crBH

Condition on independently trained, fixed scoring information. In this asymptotic analysis, calibration and test observations are iid from a fixed population, $n, m \to \infty$, $\tau_m = k/m \to \tau \in (0,1)$, and $0 < \alpha < 1 - \tau$. The ratio n/m need not approach a finite positive limit.

Define randomized population upper-tail quantiles using the independent tie labels of Section 3:

$$W_Y = 1 - F_Y(Y^-) - V^Y\{F_Y(Y) - F_Y(Y^-)\},$$
$$U_S = 1 - F_S(S^-) - U^S\{F_S(S) - F_S(S^-)\}.$$

The uniform marginals may be dependent; smaller values indicate larger augmented outcomes or scores. Even with outcome ties, population top-τ membership is $W_Y < \tau$. We use the following tail functions:

$$N_\tau(z) = \mathbb{P}(U_S \le z, W_Y \ge \tau), \qquad A_\tau(z) = z - N_\tau(z), \qquad 0 \le z \le 1.$$

These nondecreasing functions are 1-Lipschitz. Let

$$z_< = \sup\big(\{z \in [0,1] : N_\tau(z) < \alpha z\} \cup \{0\}\big),$$
$$z_\le = \sup\{z \in [0,1] : N_\tau(z) \le \alpha z\}.$$

Theorem 5 (Asymptotic FDR and power of crBH)**.** *Under the preceding iid regime, let R be the count returned by Algorithm 1, and let $P_m = TP/k$ be its realized power. For every $\eta > 0$,*

$$\mathbb{P}(R/m < z_< - \eta) \to 0, \qquad \mathbb{P}(R/m > z_\le + \eta) \to 0,$$
$$\mathbb{P}(P_m < A_\tau(z_<)/\tau - \eta) \to 0, \qquad \mathbb{P}(P_m > A_\tau(z_\le)/\tau + \eta) \to 0.$$

Consequently,

$$\frac{A_\tau(z_<)}{\tau} \le \liminf \text{Power} \le \limsup \text{Power} \le \frac{A_\tau(z_\le)}{\tau}.$$

If $z_< = z_\le =: z^* > 0$, then

$$\frac{R}{m} \xrightarrow{P} z^*, \qquad P_m \xrightarrow{P} \frac{A_\tau(z^*)}{\tau}, \qquad \text{Power} \to \frac{A_\tau(z^*)}{\tau}, \qquad \text{FDR} \to \alpha.$$

For intuition, the stable positive boundary $z_< = z_\le = z^* > 0$ holds if $g(z) = N_\tau(z) - \alpha z$ has a unique positive sign-changing crossing, with $g < 0$ on $(0, z^*)$ and $g > 0$ on $(z^*, 1]$. Theorem 5 bounds asymptotic rejection proportion and power without requiring a finite positive limit of n/m. Appendix G proves the result using uniform score-tail approximation at the adaptive cutoff. The single count/power/FDR limit needs boundary agreement; neither finite-sample FDR control nor a zero-boundary limit follows. The regime excludes fixed small k and changing trained scores.

Corollary 6 (Oracle ordering for asymptotic selection)**.** *Consider population rules $d : \mathcal{X} \to [0,1]$ with deterministic limiting selection masses and constraint $\mathbb{E}[(1 - \eta_\tau(X) - \alpha)d(X)] \le 0$, where $\eta_\tau(x) = \mathbb{P}(W_Y < \tau \mid X = x)$. An oracle power-maximizing rule is an upper level set of η_τ, with possible randomization on its boundary. A score preserving this ordering and satisfying the stable positive boundary conditions of Theorem 5 attains the oracle asymptotic power within this population rule class.*

Appendix K proves this result for the stated rule class, without asserting uniqueness or optimality over unrestricted batch-level randomization in general.

4.3 How the ranking model affects FDR and power

For crBH, the asymptotic conclusions require the boundary conditions of Theorem 5. The ranking model mainly affects power in this regime. At a fixed operating level, the relevant criterion is whether the score-induced selected region approximates the corresponding oracle upper level set of

$$\eta_\tau(x) = \mathbb{P}(W_Y < \tau \mid X = x),$$

especially near the top-k boundary. If high power is desired across operating levels, the score must preserve the ordering induced by η_τ more broadly. Distortions near the relevant top-tail boundary can assign systematically low scores to true top-k items and high scores to null items, making the projected statistics more conservative and reducing power.

Thus, the relevant notion of model quality is not global prediction accuracy but recovery of the oracle selected region at the operating cutoff, together with broader preservation of the top-tail ordering when performance across cutoffs matters. Appendix B studies this model-dependence in detail, and Subsection B.1 gives the theoretical model-class matching analysis, while the accompanying controlled experiment validates the same effect empirically.

4.4 Optional finite-sample certification

The preceding crBH construction is our primary procedure. Its guarantee is asymptotic and does not imply universal finite-sample validity. The asymptotic qualification is essential: the following small exact counterexample delineates its scope of validity.

Proposition 7 (No universal finite-sample FDR control for crBH)**.** *There is a jointly exchangeable, continuous-outcome construction with* $n = 4, m = 3, k = 1$ *and* $\alpha = 20/33$ *for which Algorithm 1 has*

$$\mathrm{FDR} = \frac{64}{105} = \frac{20}{33} + \frac{4}{1155} > \alpha.$$

Appendix H gives the exact enumeration. For settings requiring an exact finite-sample guarantee, we provide an optional conservative certification layer under exchangeability. With prespecified $\delta, \gamma > 0$ and $\delta + \gamma < \alpha$, it hardens part of the fractional reference mass using lower-rank calibration items, constructs a certified score prefix, and returns a prefix contained in the raw crBH proposal. The full risk split and construction appear in Appendix E.

Theorem 8 (Finite-sample FDR control of Certified-crBH)**.** *Under Assumption 1, with the prespecified budgets and tie handling above, Algorithm 2 satisfies* $\mathrm{FDR} \le \alpha$ *for every finite* $n, m \ge 1$ *and* $1 \le k \le m$*. In the nontrivial branch, pathwise,*

$$\widehat{S}_{\mathrm{core}} \subseteq \widehat{S}_{\mathrm{cert}} \subseteq \widehat{S}_{\mathrm{crBH}}, \qquad \mathrm{FDP}_{\mathrm{cert}} \le \mathrm{FDP}_{\mathrm{core}} + \gamma,$$

and $\mathrm{FDR}_{\mathrm{core}} \le \lambda + \epsilon_h(r_0) \le \alpha - \gamma$.

The proof is in Appendices I and J. Under a stable positive asymptotic boundary, the certification layer is first-order free (Corollary 9 in Appendix E). The main empirical analysis focuses on crBH; supplementary paired experiments quantify the conservativeness of this optional safeguard.

5 Experiments

5.1 Experimental setup

We evaluate RankNet, LambdaMART, and RankSVM on Synthetic, ESOL, Yummly28K, and MIMIC-IV (Appendix A). FDR is mean per-run FDP, power is mean true top-k recovery, and #Sel. is mean selection count. At $\alpha = 0.1$, the main benchmark uses 1,000 repetitions; sweeps use 300 with ± 1.96 MCSE, conditional on fixed application models and datasets (Appendix A.1). Beyond the main benchmark, we vary ranking quality, calibration size, target fraction, nominal level, model class, and rank projection.

5.2 Finite-sample empirical performance of crBH

High-yield settings show near-nominal empirical FDR. Table 1 gives FDR/power $0.0970/0.9339$ for Synthetic/RankNet, $0.0974/0.9586$ for Synthetic/RankSVM, and $0.0962/0.8207$ for Yummly28K/RankSVM. MIMIC-IV/RankNet gives $0.100063/0.2855$; its FDR exceeds 0.1 by only 0.000063 (0.22 MCSE), indistinguishable at this Monte Carlo resolution.

Low discovery yield accompanies conservative empirical FDR. ESOL/RankSVM selects 0.2 items on average, with power 0.0012 and FDR 0.0034; FDR must therefore be read alongside yield.

Table 1: **Uncalibrated model top-k versus** crBH **on the same main-benchmark trials.** At $\alpha = 0.1$, each configuration uses 1,000 paired repetitions. Model top-k always selects exactly k items per trial; its power equals model precision@k and its FDP equals one minus that quantity. It has no FDR guarantee. crBH is the FDR-aware selective procedure; #Sel. is its mean selection count. Yummly28K uses the artificial hidden-anchor target.

		Model top-k		crBH (ours)		
Dataset	Model	FDR	Power	FDR	Power	#Sel.
Synthetic	RankNet	0.0790	0.9210	0.0970	0.9339	518.3
	LambdaMART	0.2185	0.7815	0.0938	0.4721	263.3
	RankSVM	0.0642	0.9358	0.0974	0.9586	532.3
ESOL	RankNet	0.3165	0.6835	0.0442	0.0490	6.4
	LambdaMART	0.4303	0.5697	0.0645	0.0873	10.4
	RankSVM	0.4612	0.5389	0.0034	0.0012	0.2
Yummly28K	RankNet	0.2592	0.7408	0.0875	0.3982	45.4
	LambdaMART	0.3480	0.6520	0.0706	0.2173	25.3
	RankSVM	0.1152	0.8848	0.0962	0.8207	92.6
MIMIC-IV	RankNet	0.2721	0.7279	0.1001	0.2855	2375.6
	LambdaMART	0.3151	0.6849	0.0994	0.0521	435.9
	RankSVM	0.2760	0.7240	0.0994	0.2189	1821.3

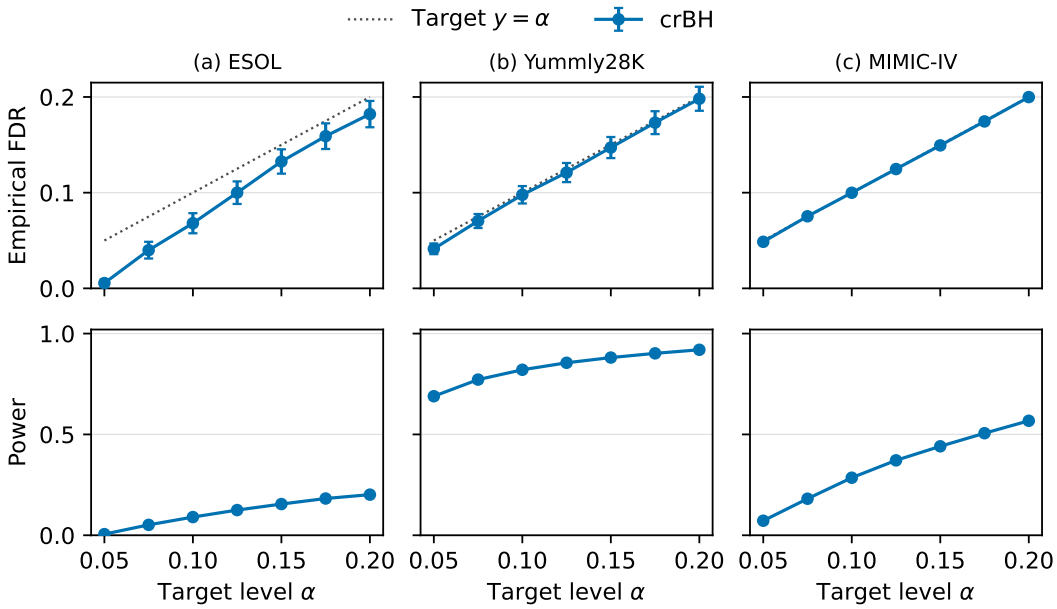

Figure 1: **Empirical FDR and power of** crBH **across nominal levels.** Columns show ESOL/LambdaMART, Yummly28K/RankSVM, and MIMIC-IV/RankNet; the top row plots mean per-run FDP against α, and the bottom row plots power. Each point uses 300 Monte Carlo repetitions with ± 1.96 MCSE error bars. The dotted diagonal $y = \alpha$ is the nominal FDR reference. Yummly28K uses the artificial hidden-anchor target.

Better ranking quality primarily raises power. Synthetic noise $0.9 \to 0.35$ raises power $0.247 \to 0.823$ with empirical FDR near 0.1 (Appendix 4). **Model-class matching changes top-tail power.** Linear-sparse signals favor RankSVM (0.728), threshold interactions LambdaMART (0.537), and hidden-anchor signals RankNet (Section B; Appendix B.2).

Small calibration budgets make selection more conservative. For Synthetic/RankSVM, $n/m : 0.05 \to 1$ raises power $0.862 \to 0.962$ and empirical FDR $0.060 \to 0.099$.

Larger k/m improves recovery. For $k/m : 0.05 \to 0.5$, power rises $0.563 \to 0.958$ with endpoint FDR $0.087/0.099$ (Appendix 4).

Nominal level trades reliability for discovery yield. At $\alpha = 0.20$, Yummly28K/MIMIC-IV give empirical FDR $0.198/0.200$ and power $0.920/0.568$; ESOL remains conservative (Figure 1).

Exact projection matters most when rank scales differ. Mass-matched step shares BH and reference mass. At $n/m = 0.05$, exact/step FDR is $0.0605/0.0782$ and power $0.8619/0.8958$; at $n = m$ both coincide. Paired Yummly28K/RankSVM gives $0.0978/0.1034$ FDR and $0.8205/0.8326$ power. Lower empirical FDR costs power (Appendix C.2, Figure 5).

5.3 Comparison with uncalibrated model top-k

Uncalibrated model top-k often exceeds target FDR. On the same 1,000 trial pairs, this rule returns k without calibration and exceeds 0.1 empirical FDR in 10/12 cases (0.1152–0.4612); Synthetic/RankNet and RankSVM are exceptions ($0.0790/0.0642$; Figure 2(a)).

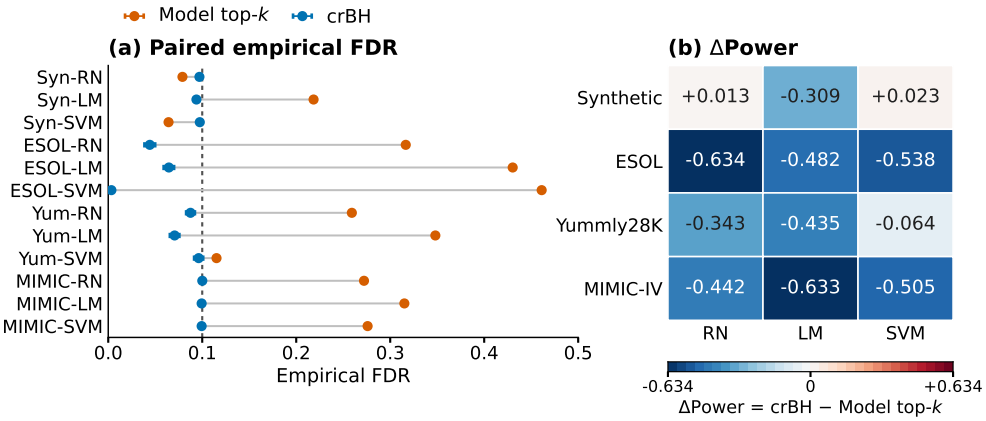

Figure 2: **Model top-k versus crBH on the same 1,000 locked trials per configuration.** (a) Paired empirical FDR with ± 1.96 MCSE bars; connectors link configurations and the dotted line marks $\alpha = 0.1$. (b) Signed power difference, crBH minus model top-k, on a zero-centered color scale. Model top-k returns k items without an FDR guarantee, so read power with panel (a). Syn/Yum/MIMIC denote Synthetic/Yummly28K/MIMIC-IV; RN/LM/SVM denote RankNet/LambdaMART/RankSVM.

FDR calibration has a model-dependent power cost. Versus unconstrained top-k (Figure 2(b)), crBH often shortens the prefix, but selects more than k with positive power differences in two strong Synthetic settings. **Exact finite-sample certification can be conservative.** The optional safeguard lowers FDR but often reduces power under its assumptions (Appendix D).

6 Limitations

The crBH FDR and power guarantee is asymptotic; finite-sample observations do not imply universal control (Proposition 7). Its optional exact certificate requires joint calibration/test exchangeability and can be conservative; shift, drift, or dependence may invalidate it and affect raw crBH. The target uses randomized ties (Section 3) and requires reliable calibration ranks. Large pools may need batched or approximate score comparisons at scale.

7 Conclusion

Exact rank projection yields crBH for inductive top-k selection, with asymptotic FDR and power characterization. Experiments show near-nominal empirical FDR in high-yield settings and how ranking quality, model class, and rank-scale mismatch shape power. An optional certificate controls finite-sample FDR under joint exchangeability.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript polishing and editing, to check mathematical derivations and proof steps, and to help write and debug experimental code. All AI-assisted theoretical arguments, code, and manuscript changes were reviewed and verified by the authors. The authors made the final research decisions and take full responsibility for all claims, proofs, experiments, code, and text in this work.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison-Wesley Professional, 2nd edition, 2011.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021. doi: 10.1145/3478535.

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.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96. ACM, 2005.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007.

W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Professional, 2009.

John S. Delaney. ESOL: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, 2004.

William Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*. John Wiley & Sons, 3 edition, 1968. ISBN 9780471257080.

Jean-Baptiste Fermanian, Pierre Humbert, and Gilles Blanchard. Transductive conformal inference for full ranking. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0212. URL https://arxiv.org/abs/2501.11384v3.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An extensive data processing pipeline for MIMIC-IV. In *Proceedings of the 2nd Machine Learning for Health Symposium*, volume 193 of *Proceedings of Machine Learning Research*, pp. 311–325, 2022.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

Anna Lena Hunkenschroer and Christoph Luetge. Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 178:977–1007, 2022.

Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, 2010.

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

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142. ACM, 2002.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Wenbo Liao, Huipeng Huang, Chen Jia, Huajun Xi, Hao Zeng, and Hongxin Wei. Distribution-informed efficient conformal prediction for full ranking. *arXiv preprint arXiv:2601.23128*, 2026.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Ariane Marandon. Conformal link prediction for false discovery rate control. *TEST*, 33(4):1062–1083, 2024. doi: 10.1007/s11749-024-00934-w.

Francesco Ricci, Lior Rokach, and Bracha Shapira (eds.). *Recommender Systems Handbook*. Springer, 2nd edition, 2015.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

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

Xiwang Yang, Harald Steck, Yang Guo, and Yong Liu. On top-k recommendation using social networks. In *Proceedings of the sixth ACM conference on Recommender systems*, RecSys 2012, pp. 67–74. ACM, September 2012. doi: 10.1145/2365952.2365969. URL http://dx.doi.org/10.1145/2365952.2365969.

Yummly28K Dataset Release. Yummly28k dataset readme. Bundled dataset documentation included with the local Yummly28K release used in this work, 2017. The readme attributes the dataset to Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz, "Being a Super Cook: Joint Food Attributes and Multi-Modal Content Modeling for Recipe Retrieval and Exploration," IEEE Transactions on Multimedia, 2017.

A DATASETS USED IN THE EXPERIMENTS

Synthetic. The synthetic benchmark is generated in-house. It is used in two roles: the main validation table, where we evaluate standard selective top-k performance under controlled exchangeable sampling, and the targeted synthetic studies, where we vary either the noise level or the geometry of the top-tail signal. This lets us isolate how score quality and model-class matching affect power without changing the crBH selection procedure. The main synthetic benchmark uses noise scale 0.2.

ESOL. ESOL (Delaney, 2004) is a small molecular-solubility benchmark with continuous labels. We treat it as a ranking problem by using the observed solubility values to define the test-pool top-k set. In the experiments, ESOL serves as a compact real-data benchmark where modest changes in ranking quality produce visible changes in selective power. The main ESOL experiment uses Morgan-only features and a 564/225/339 training/calibration/test split.

Yummly28K. Yummly28K provides real recipe covariates from the dataset release documented by Yummly28K Dataset Release (2017). In our experiment, the ranking target is semi-synthetic: a fixed hidden recipe/ingredient anchor defines an artificial ingredient-distance response. The experiment therefore evaluates top-k selection for this constructed target rather than measured user relevance or real image-query relevance. Relative to ESOL, this task provides a setting where the gap between a weak and a strong ranking model is much larger in the final selective power.

MIMIC-IV first-24h LOS. The MIMIC-IV task is a 24-hour landmark ICU resource-planning and risk-stratification problem (Johnson et al., 2023). We retain ICU stays present at 24 hours, set prediction time to ICU admission plus 24 hours, and use only information available by that landmark. The outcome is remaining ICU length of stay, defined as total ICU LOS minus the first day; within each test split, the true target set is the 50% of stays with the largest remaining LOS. Features include admission-time demographics and administrative covariates plus summaries of first-day physiology, labs, and treatments (vitals, oxygenation and FiO2, neurologic scores, labs, organ function markers, mechanical ventilation, vasopressors, dialysis, fluids, blood products, procedure counts, abnormal-value burden, within-window trends, fluid balance, and SOFA/OASIS-like severity proxies). To avoid leakage, we remove LOS and remaining-LOS fields, ICU discharge time, discharge destination, last care unit, mortality- or discharge-coded outcomes, ICD diagnosis/procedure codes, and any variable recorded after 24 hours; events must occur within the first day, stored chart values must be available by the landmark, and input amounts crossing the cutoff are prorated to the overlap. This follows the standard separation between an observation window and future ICU outcomes used in clinical time-series benchmarks and MIMIC-IV pipelines (Harutyunyan et al., 2019; Gupta et al., 2022). We evaluate selective ranking over 1,000 Monte Carlo calibration/test resamplings, aiming not to predict exact discharge day but, after the first ICU day, to flag patients likely to be in the longer half of remaining ICU stays for risk stratification and resource planning.

A.1 REPRODUCIBILITY, COMPUTE, AND ASSET NOTES

Code and data access. The supplementary paired result package records crBH and Certified-crBH on identical splits, trained scores, and tie labels. The main benchmark and each certification-budget configuration use 1,000 pairs; the synthetic, nominal-level, target-rank, and model-matching sweeps use 300 pairs per configuration. Primary budgets are $\delta/\alpha = \gamma/\alpha = 0.05$. Trial-level records, summary means and MCSE, paired differences, configuration records, and a SHA-256 manifest provide provenance for all empirical tables and figures. The synthetic RankNet code is available at `https://anonymous.4open.science/r/crBH-4D5F`; that historical demonstration is separate from the paired evaluations. ESOL uses the processed benchmark table; MIMIC-IV requires official PhysioNet credentialing and is not redistributed. Yummly28K uses real recipe covariates with the semi-synthetic target described above.

For the uncalibrated model top-k baseline, the formal runner sorts test scores in descending order with a stable sort, so score ties follow the stored test-array order; crBH instead uses its recorded randomized score-tie labels. Both use the same randomized outcome ties to define the true test top-k. The locked aggregate retains precision@k but not raw test scores or boundary tie multiplicities. Hence boundary-tie frequencies and a randomized-tie top-k recomputation cannot be recovered for all 12 configurations from these artifacts.

Compute resources. All experiments use single-machine training and evaluation; we do not use distributed training, TPU jobs, or multi-node hyperparameter sweeps. RankNet runs on standard PyTorch backends (CPU, CUDA, or Apple MPS), while LambdaMART and RankSVM are lightweight single-machine baselines.

Asset licenses and terms of use. The released synthetic supplementary code is distributed under the MIT License. ESOL is credited to the original Delaney benchmark (Delaney, 2004) and used as a small research benchmark in processed tabular form. The bundled Yummly28K readme states that the dataset was crawled through the Yummly API and should be used only for non-commercial research and/or educational purposes; we follow that restriction and use it only for benchmark evaluation. MIMIC-IV is credited to Johnson et al. (2023) and is not redistributed: access is restricted to credentialed PhysioNet users under the PhysioNet Credentialed Health Data License 1.5.0 and the PhysioNet Credentialed Health Data Use Agreement 1.5.0.

B How does model choice affect power?

B.1 Formal analysis for model-class matching

Model-class matching and the operating cutoff. Let $A^\star = \{x : \eta_\tau(x) > c\}$ and $A_S = \{x : S(x) > s\}$. If g is strictly increasing and $\|S - g \circ \eta_\tau\|_\infty \le e$, then

$$A_S \triangle A^\star \subseteq \{|g(\eta_\tau) - g(c)| \le e + |s - g(c)|\}.$$

Thus the difference in true-positive mass is at most the probability of this band, and the power difference is at most that probability divided by τ. This accounts for the actual cutoff displacement; a score approximation bound alone does not control a moving selection boundary. The oracle ordering argument is proved in Appendix K. A margin assumption can supply a rate once the cutoff displacement is controlled. Linear-sparse, threshold-interaction and metric / hidden-anchor geometries motivate the original controlled experiment below.

B.2 Controlled model-class matching experiment

Experimental analysis. Figure 3 provides a controlled validation of this theory using the same fitted ranking model and score realization within each paired trial. For raw crBH, RankSVM has the highest power in the linear-sparse regime (0.728), LambdaMART in the threshold-interaction regime (0.537), and RankNet in the metric / hidden-anchor regime (0.277). The other two models make almost no discoveries in the hidden-anchor regime. Thus the power-maximizing ranking model changes with the decision geometry. The optional certificate preserves this qualitative model ordering but reduces power at each fixed model–regime pair; the corresponding powers for the three leading models are 0.523, 0.328, and 0.087.

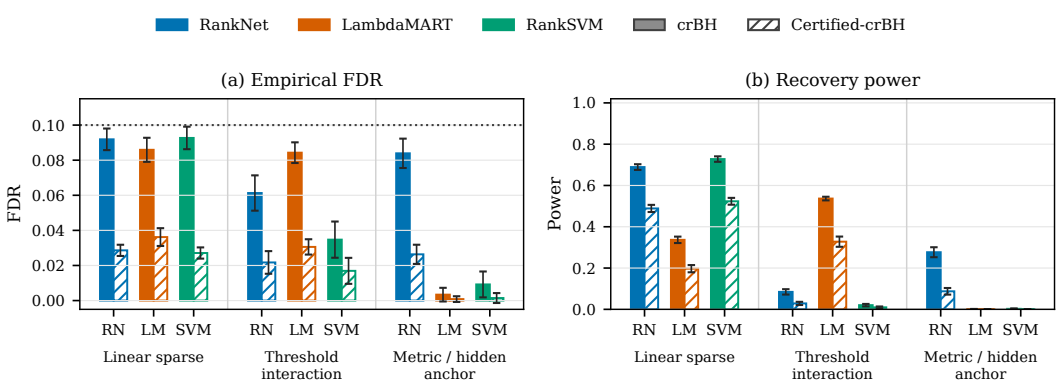

Figure 3: **Model-class matching for** crBH **power.** Paired trials compare crBH and the optional Certified-crBH supplement for RankNet (RN), LambdaMART (LM), and RankSVM (SVM) across three synthetic regimes at $\alpha = 0.1$. Filled bars denote crBH and hatched bars denote Certified-crBH, which uses $\delta/\alpha = \gamma/\alpha = 0.05$. Error bars are ± 1.96 Monte Carlo standard errors over 300 trial pairs per configuration; the dotted line is the nominal FDR target.

C Additional finite-sample displays for crBH

C.1 Controlled sensitivity to ranking quality, calibration size, and target fraction

Figure 4 gives the full crBH-only sensitivity analysis referenced in Section 5.2. Each panel retains the frozen values and Monte Carlo error bars.

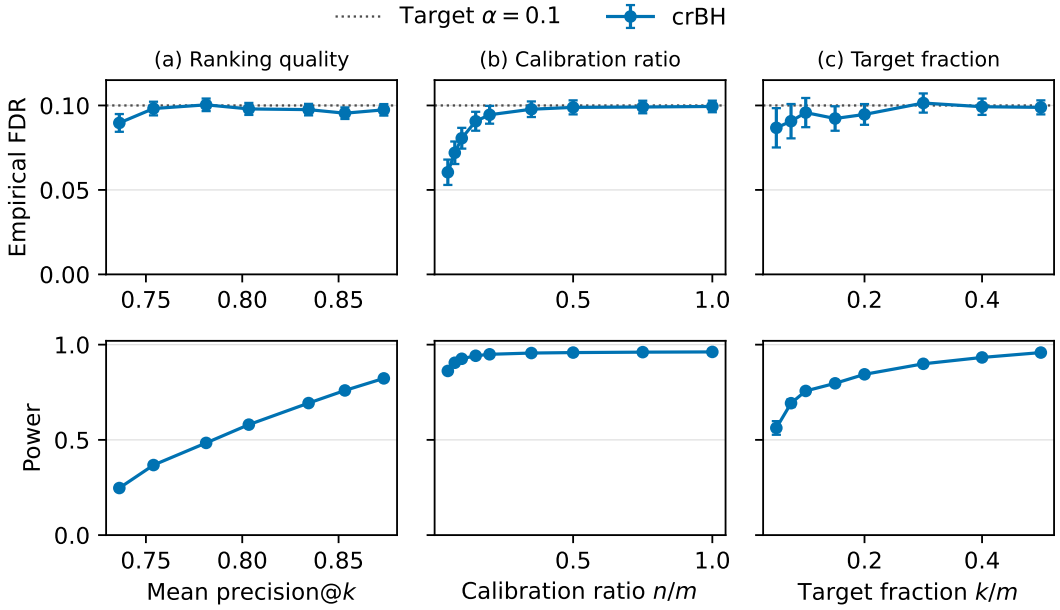

Figure 4: **Finite-sample empirical sensitivity of** crBH**.** Columns vary (a) mean precision@k, (b) calibration ratio n/m, and (c) target fraction k/m; the top row shows mean per-run FDP (empirical FDR) and the bottom row shows power. Each point uses 300 Monte Carlo repetitions. Error bars are ± 1.96 MCSE; the dotted horizontal line marks the nominal FDR level $\alpha = 0.1$. The low-n/m configurations are more conservative.

C.2 Exact rank projection versus a scaled-rank approximation

To isolate the projection shape, set $a = nk/m$ and $c = \lceil a \rceil$ and define the mass-matched scaled-rank step weights for calibration rank $r \in [n]$ by

$$q_r^{\text{step}} = \begin{cases} 0, & r < c, \\ c - a, & r = c, \\ 1, & r > c. \end{cases}$$

Their total mass is $\sum_r q_r^{\text{step}} = n - a = n(m-k)/m = \sum_r q_r^\star$. At $n = m$, both reduce to $\mathbf{1}\{r > k\}$. Exact and step use the same trained scores, calibration/test observations, outcome and score tie labels, projected-statistic formula, normalization, and π_0-corrected BH step-up rule. Thus the comparison changes only the rank-weight shape, not total null-reference mass or the critical value. Unlike uncalibrated model top-k, the step comparator uses the same FDR-aware machinery as crBH, but no finite-sample guarantee is claimed for it.

The primary diagnostic reuses the formal synthetic calibration-ratio grid. Calibration sizes are 50, 75, 100, 150, 200, 350, 500, 750, and 1000, with $m = 1000, k = 500, \alpha = 0.1$, one fixed RankSVM model, and 300 paired trials per ratio. At $n/m = 0.05$, exact versus step gives FDR 0.0605 versus 0.0782 and power 0.8619 versus 0.8958: the lower observed FDR comes with lower recovery. The paired exact-minus-step differences are -0.0178 FDR (MCSE 0.0018) and -0.0339 power (MCSE 0.0034). Both gaps shrink across this grid and are exactly zero at $n = m$, as required by the equal-size identity. Figure 5 displays the weight shapes and full ratio sweep.

Table 2 reports six secondary ESOL and Yummly28K configurations, each with one fixed model and 300 paired resamplings under the formal data, feature, split, and hyperparameter settings. Exact has lower or equal empirical FDR in all six, and lower or equal power in all six; this is a measured tradeoff, not a claim of uniform improvement. These are new paired trials, not a replay of frozen main-table predictions: the original model files were unavailable, and parameter- and split-matched caches differ from frozen model precision in ESOL/RankNet, ESOL/RankSVM, and Yummly28K/RankNet. Model training and finite-dataset uncertainty are not represented by the displayed MCSEs. The exact implementation removes at most 2.1×10^{-10} of numerical total-mass roundoff before comparing the two selectors; all paired mass discrepancies are below 10^{-10}.

Dataset	Model	Exact		Scaled step		Paired exact − step	
		FDR	Power	FDR	Power	ΔFDR	ΔPower
ESOL	RankNet	0.049	0.055	0.057	0.060	-0.008	-0.005
ESOL	LambdaMART	0.068	0.090	0.073	0.099	-0.005	-0.009
ESOL	RankSVM	0.006	0.002	0.006	0.002	0.000	0.000
Yummly28K	RankNet	0.083	0.391	0.087	0.399	-0.004	-0.008
Yummly28K	LambdaMART	0.073	0.226	0.083	0.242	-0.010	-0.016
Yummly28K	RankSVM	0.098	0.820	0.103	0.833	-0.006	-0.012

Table 2: **Public-data exact-versus-scaled-rank ablation.** Each row uses 300 paired calibration/test resamplings of one fixed trained model at $\alpha = 0.1$. Both selectors share scores, ties, total reference mass, and the step-up rule; only projection shape differs. Values are means of per-trial FDP (FDR) and power. Paired MCSEs and mean selection counts are in the machine-readable summary. The fixed models are not always identical to those behind the frozen main table; model-precision alignment is audited separately.

C.3 Visualization of the selection rule

Figure 6 illustrates how the projected statistics separate true test top-k items from other items in one ESOL/RankNet repetition. The plot illustrates the rule, not an additional validity claim.

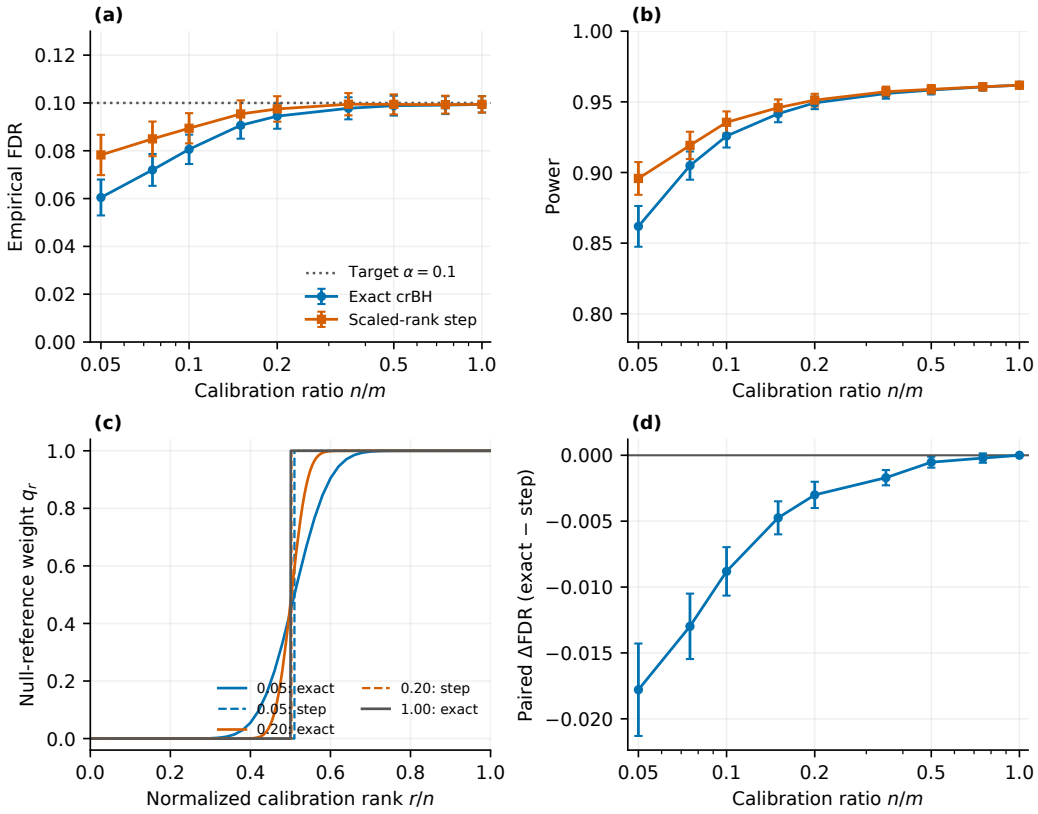

Figure 5: **Exact versus scaled-rank projection.** Synthetic calibration-ratio sweep, 300 paired trials at each of nine ratios, $m = 1000, k = 500, \alpha = 0.1$. (a) Empirical FDR with nominal target; (b) power; (c) exact and mass-matched step weights against normalized calibration rank for $n/m = 0.05, 0.20, 1.00$ (the two curves coincide at 1.00); (d) paired exact-minus-step FDR difference. Error bars are ± 1.96 MCSE, using paired differences in (d). Both methods share the same step-up rule and differ only in projection weights.

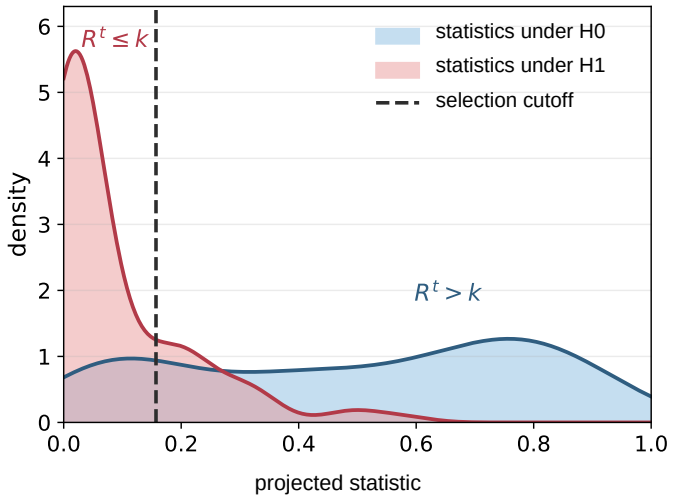

Figure 6: **Visualization of** crBH **selection rule.** Empirical projected-statistic densities from one ESOL + RankNet repetition are shown for true top-k and non-top-k items.

D SUPPLEMENTARY FINITE-SAMPLE CERTIFICATION EXPERIMENTS

These experiments are supplementary to the main empirical study of crBH. They quantify the conservativeness and power cost of replacing the asymptotic procedure by an optional exact finite-sample certificate. All comparisons below reuse paired splits, trained scores, and tie labels; the primary certification budgets are $\delta/\alpha = \gamma/\alpha = 0.05$.

Table 3: **Supplementary comparison with the optional finite-sample certificate across four datasets and three ranking models.** Each configuration uses 1,000 paired Monte Carlo runs with identical splits, trained scores, and tie labels for crBH and Certified-crBH at $\alpha = 0.1$. Certified-crBH uses the prespecified budgets $\delta/\alpha = \gamma/\alpha = 0.05$. Entries are mean FDP (FDR), mean power, and mean selection count; MCSE measure conditional Monte Carlo uncertainty. Yummly28K uses the artificial hidden-anchor target.

Dataset	Model	crBH			Certified-crBH		
		FDR	Power	#Sel.	FDR	Power	#Sel.
Synthetic	RankNet	0.0970	0.9339	518.3	0.0394	0.8413	438.6
	LambdaMART	0.0938	0.4721	263.3	0.0544	0.3498	186.9
	RankSVM	0.0974	0.9586	532.3	0.0349	0.8771	455.1
ESOL	RankNet	0.0442	0.0490	6.4	0.0113	0.0086	1.1
	LambdaMART	0.0645	0.0873	10.4	0.0239	0.0353	4.1
	RankSVM	0.0034	0.0012	0.2	0.0009	0.0003	0.1
Yummly28K	RankNet	0.0875	0.3982	45.4	0.0442	0.2838	30.4
	LambdaMART	0.0706	0.2173	25.3	0.0141	0.0557	6.1
	RankSVM	0.0962	0.8207	92.6	0.0202	0.5639	58.0
MIMIC-IV	RankNet	0.1001	0.2855	2375.6	0.0816	0.2063	1683.2
	LambdaMART	0.0994	0.0521	435.9	0.0835	0.0368	302.9
	RankSVM	0.0994	0.2189	1821.3	0.0809	0.1171	957.1

The paired main benchmark in Table 3 shows heterogeneous power costs. For example, Synthetic/RankSVM power changes from 0.9586 for crBH to 0.8771 after certification, while Yummly28K/RankSVM changes from 0.8207 to 0.5639. Figures 7–9 display the corresponding sensitivity and target-size comparisons. These empirical costs do not alter the finite-sample guarantee of Theorem 8 or establish a finite-sample guarantee for raw crBH.

D.1 CERTIFICATION-BUDGET SENSITIVITY

The primary $0.05/0.05$ setting was fixed before inspecting these sensitivity results. Figure 10 varies δ/α and γ/α independently over $\{0.025, 0.05, 0.10, 0.20\}$. Each grid cell pools 1,000 paired runs from each of the 12 dataset–model configurations, so the displayed averages weight configurations equally. These summaries describe this finite collection of tasks, not a common population estimand. Both budgets change the empirical recovery and abstention patterns; the sensitivity grid is not used to select new primary parameters.

D.2 CERTIFICATION RETENTION ACROSS DATASETS AND MODELS

We summarize retention as $R_{\text{cert}}/\max(R_{\text{raw}}, 1)$ alongside exact-equality and nonempty frequencies. Retention is zero when the raw output is empty, while exact equality then holds automatically. Thus high unconditional equality frequency need not mean that useful nonempty selections are preserved. Synthetic/RankSVM and Synthetic/RankNet have mean retention 0.8551 and 0.8464. In contrast, ESOL/RankSVM has exact equality frequency 0.993, but raw and certified nonempty frequencies are only 0.007 and 0.002, respectively; their mean counts are 0.229 and 0.059. Conditional on a nonempty raw output, the certified output is nonempty in $2/7$ of those runs. This explains why retention, abstention, and selection counts must be interpreted together.

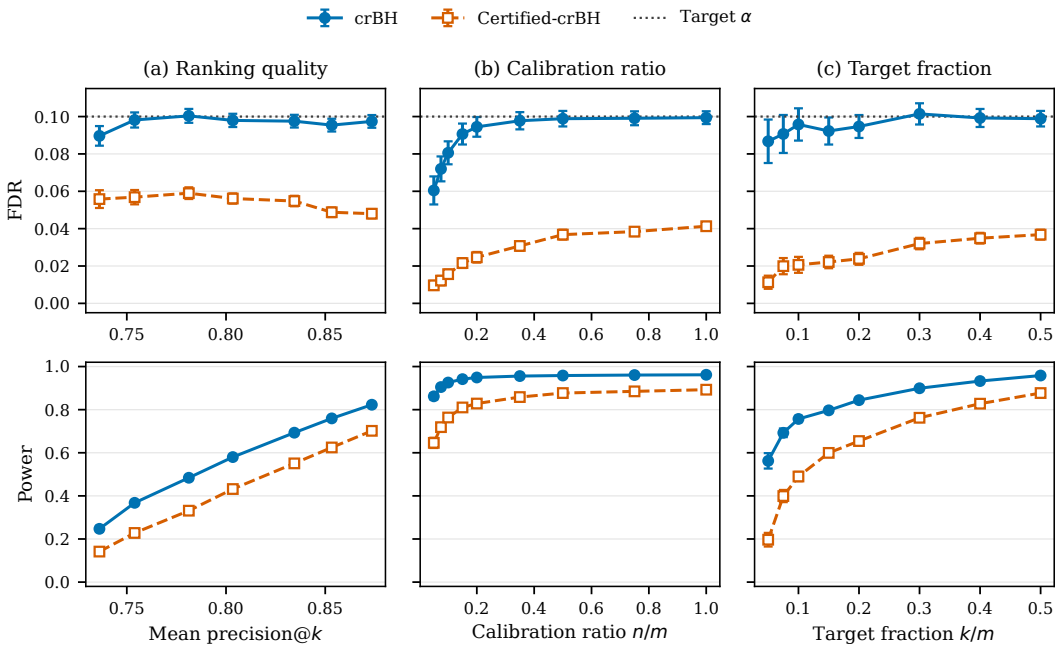

Figure 7: **Supplementary paired synthetic sensitivity.** Across ranking quality, calibration ratio n/m, and target fraction k/m, the top row shows empirical FDR and the bottom row power for crBH and Certified-crBH. Each configuration has 300 paired Monte Carlo repetitions; error bars are ± 1.96 MCSE and the dotted horizontal line is the nominal FDR level $\alpha = 0.1$. The certified procedure uses the prespecified budgets $\delta/\alpha = \gamma/\alpha = 0.05$.

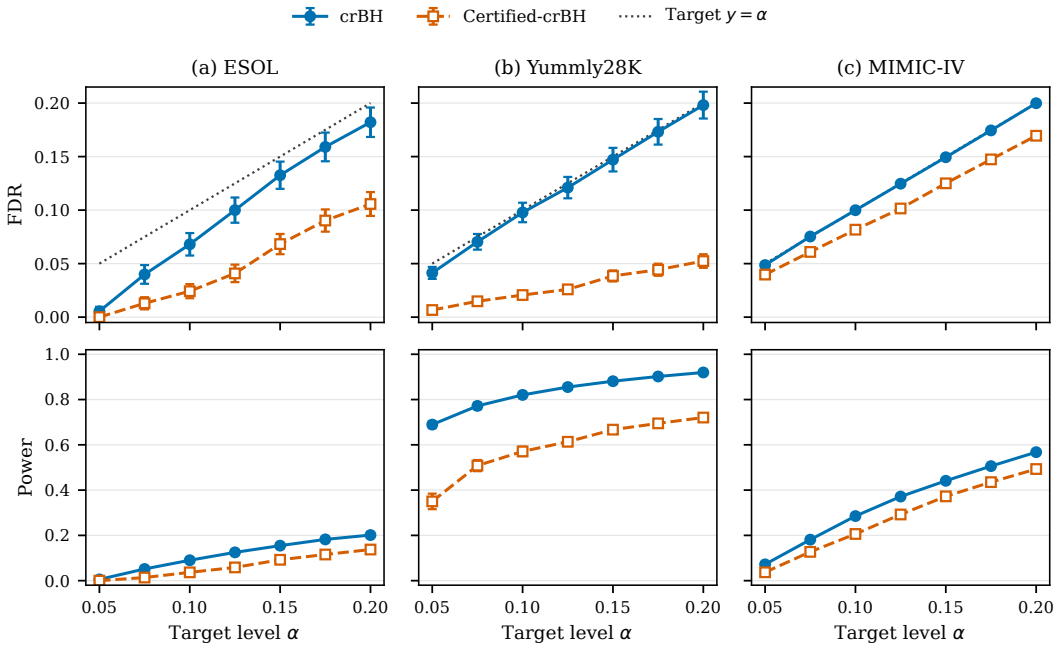

Figure 8: **Supplementary paired nominal-level sweep.** ESOL/LambdaMART, Yummly28K/RankSVM, and MIMIC-IV/RankNet show empirical FDR (top) and power (bottom) for crBH and Certified-crBH as α varies. Each configuration has 300 paired Monte Carlo repetitions with ± 1.96 MCSE error bars; the dotted diagonal is the nominal FDR reference $y = \alpha$. The certified procedure uses $\delta/\alpha = \gamma/\alpha = 0.05$.

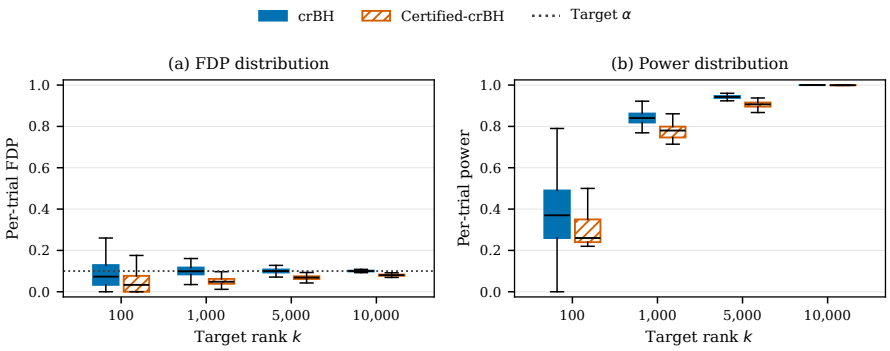

Figure 9: **Supplementary Yummly28K target-size sweep.** Paired Yummly28K/RankNet trials (300 pairs at each of four k values) show per-trial FDP and power for crBH and Certified-crBH at $\alpha = 0.1$ and $\delta/\alpha = \gamma/\alpha = 0.05$. Boxes use 1.5-IQR whiskers; hidden fliers remain in all summaries. The horizontal α line is a nominal FDR reference, not a per-trial FDP bound.

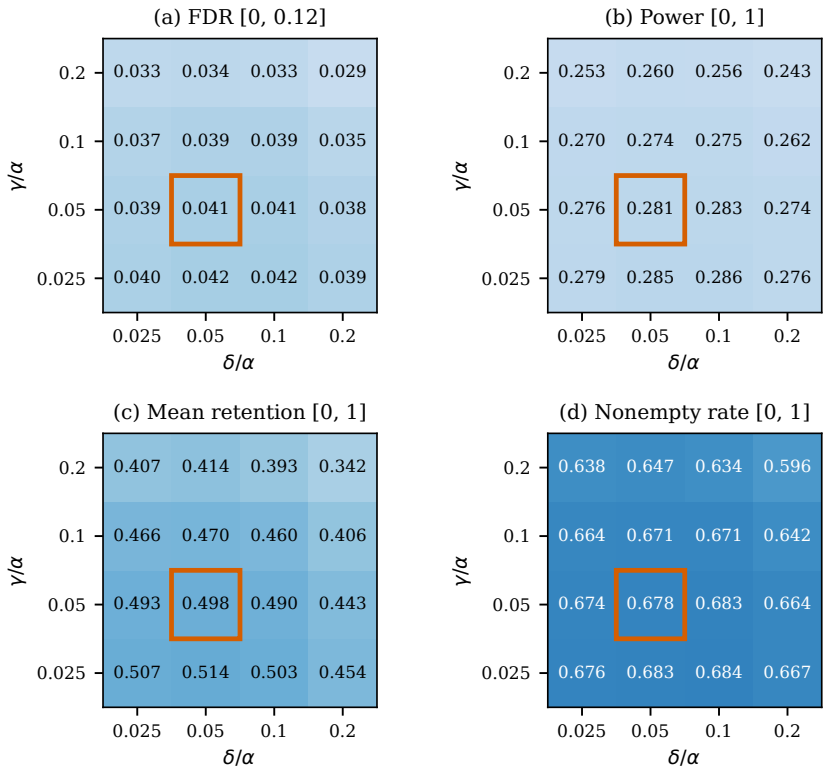

Figure 10: **Certification-budget sensitivity.** Certified-crBH empirical FDR, power, mean retention, and nonempty frequency averaged equally over the 12 main dataset–model configurations, with 1,000 paired runs per configuration and cell at $\alpha = 0.1$. Axes show δ/α and γ/α; orange outlines mark the prespecified $0.05/0.05$ primary setting. The four panels use the explicitly labelled color ranges.

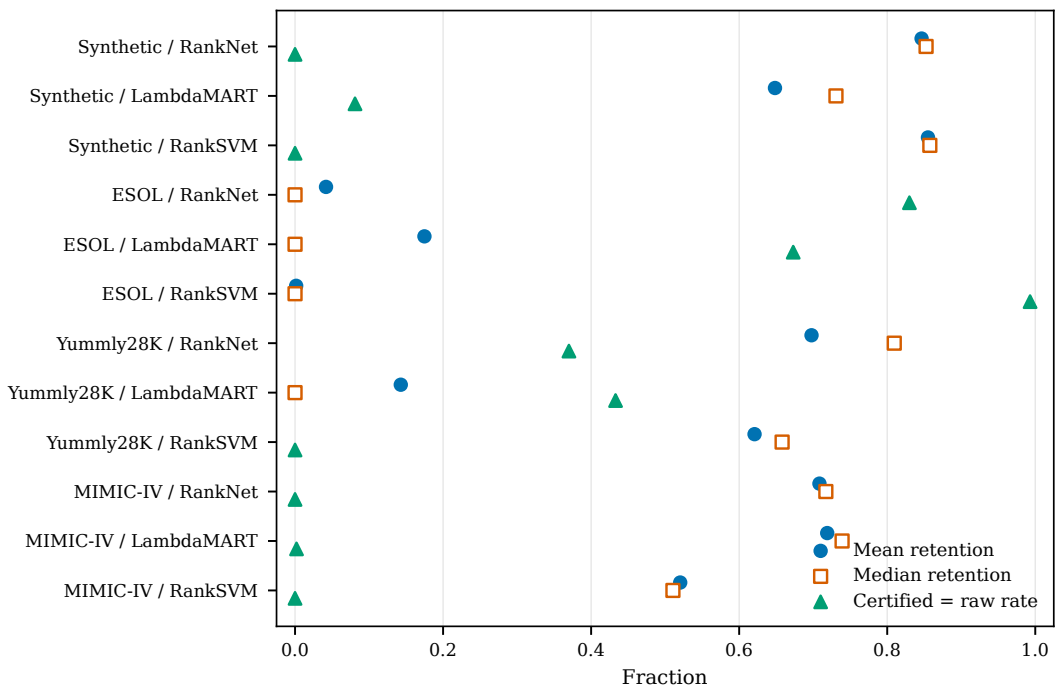

Figure 11: **Retention and equality measure different aspects of certification.** Main paired benchmark, 1,000 runs per dataset–model configuration, $\alpha = 0.1$ and $\delta/\alpha = \gamma/\alpha = 0.05$. Points show mean and median retention and the probability that certified and raw outputs coincide. Empty raw outputs contribute zero retention but count as exact equality.

E Details of the optional finite-sample certificate

If $\alpha \geq \pi = (m-k)/m$, the certificate returns all items. Otherwise prespecify $\delta, \gamma > 0$ with $\delta + \gamma < \alpha$, set $\lambda = \alpha - \delta - \gamma$ and $r_0 = \lceil m/[\lambda(n+1)] \rceil$, and return the empty set if $r_0 > m$. The budgets depend only on n, m, k, α or independent training/validation information. For $r_0 \leq m$, choose the largest h whose exact beta-binomial overshoot bound $\epsilon_h(r_0)$ is at most δ.

For the nontrivial branch $r_0 \leq m$, let $B_0 = 0$ and $B_h \sim \mathrm{BetaBinomial}(m, h, n-h+1)$ for $h = 1, \ldots, n$. The exact overshoot bound used to choose the largest feasible h is

$$\epsilon_h(r_0) = \mathbb{E} \min\left\{1, \frac{(B_h - k)_+}{r_0}\right\} = \frac{1}{r_0} \sum_{\ell=1}^{r_0} \mathbb{P}(B_h \geq k + \ell). \tag{9}$$

The zero boundary is always feasible. Form the hardened weights and projected statistics

$$w_i^{\mathrm{cert}} = \max\{q_i^\star, \mathbf{1}\{R_i^c > h\}\}, \qquad a_j^{\mathrm{core}} = \frac{1 + \sum_i w_i^{\mathrm{cert}} \mathbf{1}\{\bar{S}_i^c > \bar{S}_j^t\}}{n+1}. \tag{10}$$

For decreasing test scores $s_{(1)}, \ldots, s_{(m)}$, the hardened core step-up count is

$$\widehat{r} = \max\left(\left\{r \in \{1, \ldots, m\} : 1 + \sum_i w_i^{\mathrm{cert}} \mathbf{1}\{\bar{S}_i^c > s_{(r)}\} \leq \lambda(n+1)r/m\right\} \cup \{0\}\right). \tag{11}$$

If R_{crBH} is the raw proposal count, expand only within that proposal:

$$R_{\mathrm{cert}} = \min\left\{R_{\mathrm{crBH}}, \left\lfloor \frac{R_{\mathrm{core}}}{1-\gamma} \right\rfloor\right\}. \tag{12}$$

Algorithm 2 Certified-crBH

Require: Inputs of Algorithm 1, common score tie labels, prespecified $\delta, \gamma > 0$ with $\delta + \gamma < \alpha$ for the nontrivial branch.

$\pi \leftarrow (m-k)/m$.
if $\alpha \geq \pi$ **then**
  return all test items.
end if
$\lambda \leftarrow \alpha - \delta - \gamma$; $r_0 \leftarrow \lceil m/\lceil \lambda(n+1) \rceil \rceil$.
if $r_0 > m$ **then**
  return $\varnothing$.
end if
Compute the original $q_i^\star$ and raw proposal $\widehat{S}_{\mathrm{crBH}}$ by Algorithm 1 with the same score order.
Choose $h \leftarrow \max\{0 \leq h \leq n : \epsilon_h(r_0) \leq \delta\}$ using (9).
Form w_i^{cert} by (10) and compute $\widehat{r}$ by (11).
$R_{\mathrm{core}} \leftarrow \widehat{r}\,\mathbf{1}\{\widehat{r} \geq r_0\}$; $R_{\mathrm{crBH}} \leftarrow |\widehat{S}_{\mathrm{crBH}}|$.
$R_{\mathrm{cert}} \leftarrow \min\{R_{\mathrm{crBH}}, \lfloor R_{\mathrm{core}}/(1-\gamma) \rfloor\}$.
return the first R_{cert} test items in decreasing tie-broken score order.

Put $R_{\mathrm{core}} = \widehat{r}$ if $\widehat{r} \geq r_0$, and zero otherwise. The leading one in the numerator implies this intrinsic minimum count for a nonempty passing prefix. The core is an internal certification construction, not an additional proposed method. Algorithm 2 states the full procedure, including the all-selection and empty branches.

Corollary 9 (Certification is asymptotically first-order free)**.** *In the iid regime of Theorem 5, suppose* $\delta_{n,m}, \gamma_{n,m} \to 0$, $\delta_{n,m} + \gamma_{n,m} < \alpha$, $\log(1/\delta_{n,m}) = o(\min(n,m))$, *and* $z_< = z_\leq = z^* > 0$. *Use the intrinsic floor (which satisfies* $r_0/m \to 0$*). Then*

$$\frac{R_{\mathrm{core}}}{m}, \frac{R_{\mathrm{cert}}}{m}, \frac{R_{\mathrm{crBH}}}{m} \xrightarrow{P} z^*, \qquad \mathrm{Power}_{\mathrm{cert}} \to \frac{A_\tau(z^*)}{\tau}, \qquad \mathrm{FDR}_{\mathrm{cert}} \to \alpha.$$

This first-order statement does not assert equal finite-sample power. Its proof is in Appendix J.

F Exact rank projection

Proof of Theorem 2. For $n \leq m$, choose uniformly a size-n subset A of the m ordered slots. The rth calibration rank is mapped to the rth ordered element of A. Let $H = |A \cap \{1, \ldots, k\}|$. Then $H \sim \mathrm{Hypergeom}(m, k, n)$, and that element is outside the first k slots exactly when $H < r$. Summing its mass gives the first finite sum. In particular, $n = m$ makes $H = k$ deterministically and $q_r^\star = \mathbf{1}\{r > k\}$.

For $n > m$, merge n calibration labels and $m-1$ competitor labels in a uniform order. Conditional on a specified item's internal calibration rank r, it occupies the rth calibration slot. Its rank among itself and the competitors exceeds k exactly when at least k competitors precede it. Equivalently, at most $r-1$ calibration labels occur among the first $r+k-1$ merged positions. Their count has distribution $\mathrm{Hypergeom}(n+m-1, n, r+k-1)$, giving the second sum with the stated support.

For the mass identity in this second regime, assume $k < m$ and let H be the number of calibration labels before the kth competitor. The same event is $H < r$, and uniform arrangements give

$$\mathbb{P}(H = a) = \frac{\binom{a+k-1}{a}\binom{n-a+m-k-1}{n-a}}{\binom{n+m-1}{n}}, \quad 0 \leq a \leq n.$$

This is $\mathrm{BetaBinomial}(n, k, m-k)$, with $\mathbb{E}H = nk/m$. For $n \leq m$, the hypergeometric H also has this mean. In either case,

$$\sum_{r=1}^{n} q_r^\star = \mathbb{E}\sum_{r=1}^{n} \mathbf{1}\{H < r\} = n - \mathbb{E}H = n(m-k)/m.$$

The calibration ranks permute $1, \ldots, n$, so the same identity holds over items. If $k = m$, both projection constructions put every item inside the top-m universe and all weights are zero; the identity remains valid without invoking a beta distribution with a zero parameter.

These are probabilities under the stated auxiliary constructions, conditional on internal rank, not posterior null probabilities conditional on calibration scores or the entire observed sample. In particular the whole-pool augmentation used for $n \le m$ is not the replacement-competitor construction used for $n > m$. □

G Asymptotic FDR and power proofs

Lemma 10 (Average projection error)**.** *If $H \in \{0, \dots, n\}$ and $q_r = \mathbb{P}(H < r)$, then for every integer $a \in \{0, \dots, n\}$,*

$$\sum_{r=1}^{n} |q_r - \mathbf{1}\{r > a\}| = \mathbb{E}|H - a|.$$

For $m > 1$ and $a = \lfloor n\tau_m \rfloor$, the original piecewise weights therefore satisfy

$$\frac{1}{n}\sum_r |q_r^\star - \mathbf{1}\{r > a\}| \le b_{n,m}, \qquad b_{n,m}^2 = \frac{1}{n^2} + \begin{cases} \tau_m(1-\tau_m)\dfrac{m-n}{n(m-1)}, & n \le m,\ m > 1, \\ \tau_m(1-\tau_m)\dfrac{n+m}{n(m+1)}, & n > m. \end{cases}$$

In the asymptotic regime $b_{n,m} \to 0$, without a restriction on n/m.

Proof. For fixed H, the number of indices $r \le a$ with $H < r$, plus the number $r > a$ with $H \ge r$, is $|H - a|$. Taking expectations proves the identity. Apply $\mathbb{E}|H - a| \le \sqrt{\mathrm{Var}(H) + (\mathbb{E}H - a)^2}$ to the two distributions in the proof of Theorem 2. Their means equal $n\tau_m$ and their variances are respectively $n\tau_m(1-\tau_m)(m-n)/(m-1)$ and $n\tau_m(1-\tau_m)(n+m)/(m+1)$. The squared mean discrepancy is at most one. The first normalized variance is at most $1/(4n)$ and the second at most $(1/n + 1/(m+1))/4$. □

Proof of Theorem 5. Population labels. Write $Z_{i,m} = \mathbf{1}\{W_{Y,i}^c \ge \tau_m\}$ and $Z_i = \mathbf{1}\{W_{Y,i}^c \ge \tau\}$. Both $\{R_i^c > a\}$ and $\{Z_{i,m} = 1\}$ are lower sets in the same outcome order, so their symmetric difference has size $|n - a - \sum_i Z_{i,m}|$. The sum is binomial with parameters $n, 1 - \tau_m$. Thus

$$\mathbb{E}\frac{1}{n}\sum_i |q_i^\star - Z_i| \le b_{n,m} + \sqrt{\tau_m(1-\tau_m)/n} + n^{-1} + |\tau_m - \tau| \to 0. \tag{13}$$

The last term follows from the uniform outcome transform even at atoms. This does not assume independence of the internal ranks or of the weights. Markov's inequality gives convergence of the average error in probability.

Uniform tails. Put $Q_n(z) = n^{-1}\sum_i q_i^\star \mathbf{1}\{U_{S,i}^c < z\}$. Its uniform distance to the empirical null sub-CDF is at most the error in (13). That empirical sub-CDF converges uniformly to N_τ: on a fixed grid j/J, apply the law of large numbers and a finite union bound; between grid points monotonicity and the 1-Lipschitz population sub-CDF bound the error by the grid error plus $1/J$. First choose J large, then increase sample sizes. The same argument applies to the test score CDF and its null and alternative sub-CDFs. Strict versus weak empirical comparisons differ by at most $1/n$ or $1/m$, since the randomized quantiles have no ties. Hence

$$\sup_z |Q_n(z) - N_\tau(z)| = o_P(1), \qquad \sup_{1 \le r \le m} |U_{(r)}^t - r/m| = o_P(1).$$

The second assertion follows by evaluating the test empirical CDF at its order statistics. Since N_τ is Lipschitz, at all candidate prefixes simultaneously,

$$\sup_{1 \le r \le m} \left| n^{-1}\left[1 + \sum_i q_i^\star \mathbf{1}\{\bar{S}_i^c > s_{(r)}\}\right] - N_\tau(r/m)\right| = o_P(1). \tag{14}$$

The leading constant contributes exactly $1/n$ and has not been removed from the algorithm.

Pool-dependent target. The true test-null set and $\{j : W_{Y,j}^t \ge \tau\}$ are lower sets of sizes $m - k$ and $\sum_j \mathbf{1}\{W_{Y,j}^t \ge \tau\}$. Their symmetric difference, divided by m, converges in probability to zero by

the law of large numbers and $\tau_m \to \tau$. Combining this with the uniform empirical sub-CDF bounds and the order-statistic bound proves

$$\sup_{0\le r\le m} |V(r)/m - N_\tau(r/m)| = o_P(1), \qquad \sup_{0\le r\le m} |TP(r)/m - A_\tau(r/m)| = o_P(1). \tag{15}$$

This explicitly controls the empirical top-k target, rather than identifying it with a population null.

Step-up brackets. Eventually $\alpha < \pi_m = 1 - \tau_m$. By the exact mass identity the raw critical value divided by n is $\alpha(1+1/(n\pi_m))r/m$, which converges uniformly to $\alpha r/m$. For $g(z) = N_\tau(z) - \alpha z$, any strict feasible point $g(z) < 0$ has a nearby grid prefix $r_m/m \to z$ that passes with probability tending to one, by (14). If $z_< - \eta > 0$, choose such a point above $z_< - \eta$; otherwise the lower bracket is automatic. On the compact interval $[z_\le + \eta, 1]$, when nonempty, g has a strictly positive minimum. The uniform approximation excludes every passing prefix there with probability tending to one. This proves the upper bracket. Monotonicity and continuity of A_τ, (15), and $k/m \to \tau$ give the realized-power brackets. Boundedness of TP/k gives the expectation brackets.

Stable positive boundary. If $z_< = z_\le = z^* > 0$, the brackets give $R/m \to_P z^*$. Since $N_\tau(1) = 1 - \tau > \alpha$, $z^* < 1$. Continuity and maximality of the feasible boundary imply $N_\tau(z^*) = \alpha z^*$: a strict inequality would extend feasibility to its right. Equation (15) implies $V/m \to_P N_\tau(z^*)$ and $TP/m \to_P A_\tau(z^*)$. Division is valid on an event of probability tending to one on which $R/m \ge z^*/2$. Therefore FDP converges in probability to α and realized power to $A_\tau(z^*)/\tau$. Both are bounded by one, yielding the expected FDR and power limits. No claim is made by dividing at a vanishing boundary. □

H Exact finite-sample counterexample

Proof of Proposition 7

Proof. Order seven underlying units by decreasing outcome and give them fixed scores

$$(1, 4, 5, 6, 2, 7, 3).$$

Apply a uniform random permutation before assigning four calibration and three test positions. Independent continuous perturbations of size less than $1/4$ can be added to outcome levels separated by one, so the outcomes are continuous and their ordering does not change. The construction is jointly exchangeable. The crBH weights are $(1/3, 3/5, 4/5, 14/15)$ and its BH level is $10/11$. Among the $\binom{7}{4} = 35$ calibration subsets, the pairs (R, V) have counts $7, 2, 6, 20$ for $(0,0), (1,1), (2,2), (3,2)$, respectively. Thus

$$\mathrm{FDR} = \frac{2 + 6 + 20(2/3)}{35} = \frac{64}{105} = \frac{20}{33} + \frac{4}{1155}.$$

The accompanying verifier reproduces this calculation with rational arithmetic. This counterexample concerns the crBH rule under joint exchangeability; it does not assert failure at $\alpha = 0.1$ or for every i.i.d. distribution. □

I Finite-Sample Certification Proofs

I.1 Certification boundary rank distribution

For $h \ge 1$, let θ_h be the hth largest augmented calibration outcome and let $B_h = \#\{j : \widetilde{Y}_j^t > \theta_h\}$. Set $\theta_0 = +\infty$ and $B_0 = 0$. These numerical boundary values are proof objects, not required algorithm inputs. All comparisons use the strict augmented outcome order from Section 3.

Lemma 11 (Certification boundary exceedance distribution)**.** *For $h \ge 1$,*

$$B_h \sim \mathrm{BetaBinomial}(m, h, n - h + 1), \quad \mathbb{P}(B_h = b) = \frac{\binom{h+b-1}{b}\binom{n-h+m-b}{m-b}}{\binom{n+m}{m}}, \quad b = 0, \dots, m.$$

For $h = 0$, $B_h = 0$. Moreover,

$$\mathbb{E}B_h = \frac{mh}{n+1}, \qquad \mathrm{Var}(B_h) = mp_h(1-p_h)\frac{n+m+1}{n+2}, \quad p_h = \frac{h}{n+1}.$$

Proof. Conditional on the unordered pooled observations, their calibration/test labels are uniformly distributed in the decreasing outcome order. For $B_h = b$, the hth calibration label occupies position $h+b$. Choose the b preceding test positions and then the remaining test positions. This gives the displayed mass. Its moments are the beta-binomial moments, or follow by the equivalent representation $P \sim \mathrm{Beta}(h, n-h+1)$ and $B_h \mid P \sim \mathrm{Binomial}(m, P)$. □

I.2 Step-up and boundary control

The following lemma makes the dependence argument explicit. Its leave-one-out construction is related to the exchangeability proof in Jin & Candès (2023); we give the proof needed here, including the additional reference term and the random-boundary conditioning.

Lemma 12 (A direct step-up inequality)**.** *Consider l null reference scores C and m test scores $T_1, \dots, T_m$. Conditional on a side environment, let the null test index set $\mathcal{H}$ have fixed size M. For every $j \in \mathcal{H}$, suppose C and T_j are exchangeable conditional on the other test scores and that environment. Let u be a fixed nonnegative, nonincreasing function and $D > 0$ a fixed constant, invariant under these null-reference permutations. Define*

$$a_j = \frac{1 + \sum_{z \in C} \mathbf{1}\{z > T_j\} + u(T_j)}{D}.$$

With independent tie-breaking, the BH step-up rule at level λ obeys

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{H}|}{|\widehat{S}| \vee 1} \,\middle|\, \textit{side environment} \right] \le \lambda \frac{D}{l+1} \frac{M}{m}.$$

No PRDS or marginal p-value validity condition is asserted or needed.

Proof. Let $R = |\widehat{S}|$ be the selection count of the unmodified BH step-up rule defined in this lemma. Fix $j \in \mathcal{H}$. Set coordinate j to zero and, for $\ell \neq j$, define

$$a_\ell^{(j)} = \frac{\sum_{z \in C \cup \{T_j\}} \mathbf{1}\{z > T_\ell\} + u(T_\ell)}{D}.$$

Let $R_j \ge 1$ be the BH count for this modified vector. On the event that j is selected by the unmodified BH step-up rule defined in this lemma, a coordinate with $T_\ell < T_j$ is unchanged, whereas one with $T_\ell > T_j$ is reduced by $1/D$. In the latter case monotonicity gives $a_\ell \le a_j$, so this coordinate was already selected. Lowering only selected coordinates does not alter a BH count: all order statistics beyond the original count are unchanged. Therefore $R = R_j$ whenever j is selected by the unmodified rule.

Condition further on the unordered scores $C \cup \{T_j\}$ and the other test scores. The modified vector, and hence R_j, is invariant and fixed. By exchangeability,

$$K_j = 1 + \sum_{z \in C} \mathbf{1}\{z > T_j\}$$

is uniform on $\{1, \dots, l+1\}$, while $a_j \ge K_j/D$. Consequently,

$$\begin{aligned} \mathbb{E}\left[\frac{\mathbf{1}\{j \in \widehat{S}\}}{R \vee 1} \,\middle|\, \text{invariant environment} \right] &\le \frac{1}{R_j} \mathbb{P}\left(K_j \le \frac{\lambda D R_j}{m} \,\middle|\, \text{invariant environment} \right) \\ &\le \frac{\lambda D}{m(l+1)}. \end{aligned}$$

The last inequality also holds when the threshold exceeds $l+1$. Summing over $j \in \mathcal{H}$ and removing the additional conditioning proves the claim. Exact equality with a BH critical value creates no difficulty. □

Theorem 13 (Certification-core FDR bound)**.** *Under Assumption 1, fix $h \in \{0, \dots, n\}$, $0 < \lambda < 1$ and $r_0 \in \{1, \dots, m\}$ independently of the realized calibration/test scores, or as deterministic functions of n, m, k, α. Assign arbitrary fixed weights $q_1, \dots, q_h \in [0,1]$ to the top h calibration ranks and weight one to every lower calibration rank. Use (11) and the minimum-size rule. Then*

$$\mathrm{FDR}_{\text{top-k}} \le \lambda + \epsilon_h(r_0). \tag{16}$$

In particular, the certification core with $\epsilon_h(r_0) \le \delta$ has FDR at most $\lambda + \delta$.

Proof. First consider the boundary null set $\mathcal{H}_h = \{j : Y_j^t \le \theta_h\}$. Conditional on the unordered pooled observations, the identity of the boundary, the identities and ranks of all upper calibration units (including the boundary), the calibration/test allocation of units above the boundary, and the test index positions belonging to $\mathcal{H}_h$, the below-boundary units are still uniformly allocated to $n-h$ calibration and $m-B_h$ test positions. For $h=0$, the upper environment is empty and all calibration units are references. For $h=n$, there are zero lower calibration references, and the lemma still applies with $l=0$. Swapping a below-boundary test unit with a below-boundary calibration unit preserves the conditioned environment. The h upper calibration points and their rank weights remain fixed. They contribute the function

$$u(s) = \sum_{R_i^c \le h} q_{R_i^c} \mathbf{1}\{\bar{S}_i^c > s\},$$

which is nonnegative and nonincreasing. Apply Lemma 12 with $l = n-h$, $M = m - B_h$, and $D = n+1$ to obtain the conditional boundary-FDR bound

$$\lambda \frac{n+1}{n-h+1} \frac{m-B_h}{m}.$$

By Lemma 11, its expectation is exactly λ. Replacing an output of size less than r_0 by the empty set only decreases boundary FDP. Hence $\mathrm{FDR}_{\theta_h} \le \lambda$ after the size rule as well.

If $B_h \le k$, every test-pool top-k null is a boundary null. If $B_h > k$, there are precisely $B_h - k$ additional possible nulls above the boundary. Every nonempty output has size at least r_0, so pathwise

$$\mathrm{FDP}_{\text{top-k}} \le \mathrm{FDP}_{\theta_h} + \min\left\{1, \frac{(B_h-k)_+}{r_0}\right\}.$$

Taking expectations proves (16). No independence between B_h and the selection set is used. □

Remark 14 (The role of tuning)**.** The score model and tuning may be chosen using independent training data, provided joint calibration/test exchangeability holds conditional on that information. Selecting h, δ, or r_0 to maximize discoveries on the current calibration/test split is not covered. The largest feasible h in (9) is a deterministic, distribution-free computation, and requires no such data-dependent optimization.

J From the certification core to Certified-crBH

Proof of Theorem 8. If $\alpha \ge \pi$, selecting all items has FDP exactly $\pi \le \alpha$ for every sample. Otherwise $0 < \pi < 1$, $0 < \gamma < 1$, and $\lambda = \alpha - \delta - \gamma > 0$. If $r_0 > m$, both the core and final output are empty and FDR is zero.

Suppose $r_0 \le m$. At calibration ranks above the boundary the certification weights are the deterministic values $q_r^\star \in [0,1]$, and at ranks below the boundary they equal one. Thus Theorem 13 applies directly to this core, with the same h, λ, r_0, giving

$$\mathrm{FDR}_{\text{core}} \le \lambda + \epsilon_h(r_0) \le \lambda + \delta = \alpha - \gamma.$$

Here the numerical boundary is only a proof object; the algorithm uses internal ranks and the exactly computed rank distribution. The choice of h is independent of the realized observations.

For completeness, the subset argument uses the actual two normalizations. Since $\sum_i q_i^\star = n\pi$ and $\alpha < \pi$, the raw BH critical value for its unnormalized numerator is

$$(1+n\pi)\frac{\alpha}{\pi}\frac{r}{m} = \alpha\left(n + \frac{1}{\pi}\right)\frac{r}{m}.$$

The core numerator is at least the raw numerator because $w_i^{\text{cert}} \ge q_i^\star$. Also

$$\alpha\left(n+\frac{1}{\pi}\right) - \lambda(n+1) = (\delta+\gamma)(n+1) + \alpha(1/\pi - 1) > 0.$$

Therefore every passing core prefix passes the raw inequality. Taking maxima over all prefixes and then applying the core floor gives $R_{\text{core}} \le R_{\text{crBH}}$. Both sets are prefixes of the same tie-broken score order, so this is set inclusion, not merely a count comparison.

Because R_{core} is an integer and $0 < \gamma < 1$, $\lfloor R_{\text{core}}/(1-\gamma) \rfloor \geq R_{\text{core}}$. Consequently (12) yields

$$R_{\text{core}} \leq R_{\text{cert}} \leq R_{\text{crBH}}, \qquad R_{\text{cert}} \leq R_{\text{core}}/(1-\gamma).$$

For $R_{\text{core}} = 0$ it gives $R_{\text{cert}} = 0$ and both FDPs are zero. If $R_{\text{core}} > 0$, the added prefix segment contains at most $R_{\text{cert}} - R_{\text{core}}$ nulls. Hence

$$\begin{aligned}\text{FDP}_{\text{cert}} = \frac{V_{\text{cert}}}{R_{\text{cert}}} &\leq \frac{V_{\text{core}}}{R_{\text{cert}}} + \frac{R_{\text{cert}} - R_{\text{core}}}{R_{\text{cert}}} \\ &\leq \frac{V_{\text{core}}}{R_{\text{core}}} + \gamma = \text{FDP}_{\text{core}} + \gamma.\end{aligned}$$

Taking expectations proves $\text{FDR}_{\text{cert}} \leq \text{FDR}_{\text{core}} + \gamma \leq \alpha$. No independence of the core and its expansion, or of the core and its proposal, is required. □

Proof of Corollary 9. First $\lambda_{n,m} \to \alpha > 0$ and the intrinsic floor satisfies $r_0/m \leq 1/[\lambda_{n,m}(n+1)] + 1/m \to 0$, so eventually $r_0 \leq m$. We show $h/n \to \tau$. Fix $\eta > 0$ sufficiently small and put $h_- = \lfloor (n+1)(\tau_m - 3\eta) \rfloor$. For sufficiently large sizes $1 \leq h_- \leq n$. In the beta-binomial mixture $P \sim \text{Beta}(h_-, n - h_- + 1)$, $B \mid P \sim \text{Binomial}(m, P)$, the order-statistic identity gives

$$\mathbb{P}(P > \tau_m - \eta) = \mathbb{P}\{\text{Binomial}(n, \tau_m - \eta) \leq h_- - 1\} \leq e^{-c_\eta n}.$$

Conditional on $P \leq \tau_m - \eta$, the upper binomial tail at $k = m\tau_m$ is at most $e^{-c_\eta m}$, decreasing the positive constant if necessary. These bounds follow from exponential Markov inequality and the Bernoulli moment generating function, optimized to give $\exp[-N\,\text{kl}(x\|p)]$; the fixed mean gaps keep the divergence positive. Thus

$$\epsilon_{h_-}(r_0) \leq \mathbb{P}(B > k) \leq e^{-c_\eta n} + e^{-c_\eta m} < \delta_{n,m}$$

eventually, by the slow-budget condition. The largest feasible boundary is at least h_-. Conversely, for every feasible h and $r_0 \leq m$,

$$\delta_{n,m} \geq \epsilon_h(r_0) \geq \frac{\mathbb{E}(B_h - k)_+}{m} \geq \frac{h}{n+1} - \tau_m.$$

Together these bounds imply $h/n \to \tau$.

Let $H_i = \mathbf{1}\{R_i^c > h\}$ and $Z_i = \mathbf{1}\{W_{Y,i}^c \geq \tau\}$. Their lower-set symmetric difference has size $|n - h - \sum_i Z_i|$, so $n^{-1}\sum_i |H_i - Z_i| \to_P 0$. The original projection weights obey (13). Pointwise, for $q \in [0,1]$ and binary H, Z,

$$|\max(q, H) - Z| \leq |q - Z| + |H - Z|.$$

Therefore the certification weights converge in average absolute error to the same population null labels. Repeating the uniform-tail argument of Theorem 5 gives

$$\sup_r \left| \frac{1 + \sum_i w_i^{\text{cert}} \mathbf{1}\{\bar{S}_i^c > s_{(r)}\}}{n+1} - N_\tau(r/m) \right| \to_P 0.$$

The difference introduced by the $n+1$ normalization is at most $1/(n+1)$. The core critical value is $\lambda_{n,m} r/m \to \alpha r/m$ uniformly, and its relative floor vanishes. The same strict/weak feasible-boundary argument proves $R_{\text{core}}/m \to_P z^*$. Theorem 5 proves $R_{\text{crBH}}/m \to_P z^*$, and the pathwise inclusion in Theorem 8 sandwiches R_{cert}/m between them. Finally the uniform true-null and true-positive approximation (15) applies to every score prefix, including this certified one. Since $z^* > 0$, the same division and boundedness argument yields

$$\text{FDR}_{\text{cert}} \to N_\tau(z^*)/z^* = \alpha, \qquad \text{Power}_{\text{cert}} \to A_\tau(z^*)/\tau.$$

Thus the equality limit follows from the positive stable boundary and target approximation, not merely from the finite-sample upper bound. □

K Oracle Ordering and Model Approximation

Put $\eta_\tau(x) = \mathbb{P}(W_Y < \tau \mid X = x)$. Consider population rules $d : \mathcal{X} \to [0, 1]$ with deterministic limiting selection masses, and solve

$$\max_d \mathbb{E}[\eta_\tau(X)d(X)] \quad \text{subject to} \quad \mathbb{E}[(1 - \eta_\tau(X) - \alpha)d(X)] \leq 0.$$

This class does not include unrestricted batch-level randomization of an entire selection procedure under an average-FDR constraint.

Proposition 15 (An oracle upper-level rule exists)**.** *An optimal rule for this population problem is an upper level set of η_τ, with possible randomization on its threshold boundary.*

Proof. For each fixed rejection mass $r \in [0, 1]$, thresholding the largest values of η_τ maximizes $T(r) = \mathbb{E}[\eta_\tau(X)d(X)]$ among all rules with $\mathbb{E}d = r$. To see this, compare any such rule to an upper-quantile rule d_r: for its threshold c, $(\eta_\tau - c)(d_r - d) \geq 0$ pointwise, and the term involving c integrates to zero. Therefore $T(r)$ is the integral of the largest r fraction of posterior probabilities. It is continuous and nondecreasing. Feasibility is $T(r) \geq (1 - \alpha)r$, a closed condition on the compact interval $[0, 1]$. A maximizing feasible r exists, and its upper-level rule is optimal. This proof also covers ties, nonunique optima and limiting multipliers without assuming a finite active Lagrange multiplier. □

A score preserving the ordering of η_τ is a sufficient oracle ordering. Combined with Theorem 5, a positive stable boundary gives oracle asymptotic power within the specified population rule class at the α frontier. No universal uniqueness or "if and only if" claim about selected sets is needed.

For model approximation, suppose g is strictly increasing and $\|S - g \circ \eta_\tau\|_\infty \leq e$. If the actual score cutoff is s and the oracle posterior cutoff is c, then

$$\{S > s\} \triangle \{\eta_\tau > c\} \subseteq \{|g(\eta_\tau) - g(c)| \leq e + |s - g(c)|\}.$$

The absolute difference in true-positive mass is bounded by the probability of this band, and the power difference by that probability divided by τ. A margin assumption can turn this into a rate, but only after accounting for the cutoff displacement. Such assumptions are optional power refinements, not hypotheses of the finite-sample FDR theorem.