# CONFORMAL CONFIDENCE SUPPRESSION FOR OFFLINE AND ONLINE LEARNING-TO-RANK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Ranking systems often decide which candidates remain eligible before assigning display positions. Suppressing a useful candidate at this stage can prevent its recovery by the downstream ranker. We propose conformal confidence suppression (CoCo), a model-agnostic gate that suppresses only when a calibrated upper bound on incremental utility is negative. We distinguish eligibility from actual display, construct global, groupwise, and locally scaled gates, and prove finite-sample false-suppression control for an exchangeable calibration target. We also develop gated policy-value estimators and an online predict–gate–rank–update procedure, with conditional calibration results and a suppression-aware regret decomposition. On a public learning-to-rank benchmark, CoCo-M improves utility over the same ranker while empirical target violation tracks the chosen risk level. A sequential full-information simulation shows gains for evolving rankers, and randomized uplift data provide complementary calibration evidence. The results characterize how calibration and gate granularity trade off utility and conservative suppression.

## 1 INTRODUCTION

Learning-to-rank (LTR) is usually formulated as an ordering problem: given a request context and a candidate set, learn a scoring function that places better candidates above worse candidates. This formulation is central to search, recommendation, and advertising, and has led to a large literature on supervised, counterfactual, and online ranking methods (Liu, 2009; Li, 2011; Joachims et al., 2017; Agarwal et al., 2019; Li et al., 2019). In deployed systems, however, ranking is not the only decision. Before a downstream ranker assigns positions, the system may also decide whether a result, product, advertisement, widget, carousel, or module should remain eligible to be shown. Suppressing a useful candidate at this stage prevents the downstream ranker from displaying it. A conservative eligibility rule therefore requires reliable evidence of reduced utility before suppressing a candidate.

Eligibility decisions are particularly relevant to whole-page and marketplace ranking, where organic and sponsored content, duplicated items, and different page regions compete for limited user attention (Sorokina & Cantu-Paz, 2016; Qin et al., 2024). A candidate can have a high relevance score and still be inappropriate to show in a particular context. For example, near-duplicate organic and advertising results can crowd each other out, top-page regions may prioritize direct relevance while lower regions may support exploration, and some content can increase distraction, abandonment, or direct display cost. These interactions motivate a calibrated suppression rule that retains candidates under uncertainty and removes them only when the evidence indicates harm.

We formulate this decision as a suppression gate for LTR systems. For a context-candidate pair, $h = 0$ means suppression and $h = 1$ means impression, or display eligibility. Impression is distinct from actual display: a candidate that passes the gate may still rank below the displayed slots. We propose conformal confidence suppression (CoCo), a calibrated layer for this gate. CoCo uses a model that estimates the incremental net value of keeping a candidate eligible rather than suppressing it. A held-out calibration set then converts this estimate into a conservative one-sided decision rule. The gate suppresses a candidate only when its calibrated upper bound is negative. Otherwise, the candidate remains eligible for placement by the downstream ranker.

Our contributions are as follows. First, we formulate impression eligibility as a conservative suppression problem for LTR, explicitly distinguishing gate-level impression from realized display.



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.

Second, we introduce CoCo-1, CoCo-M, and CoCo-F, three conformal gates that provide finite-sample false-suppression control with global, groupwise, and locally scaled calibration. Third, we give offline value estimators for gated ranking policies and an online version with an explicit update loop and conditional rolling-calibration guarantees, connecting CoCo to counterfactual LTR, safe policy learning, and online ranking (Joachims et al., 2017; Agarwal et al., 2019; Gupta et al., 2024; Zhuang & Zuccon, 2020; Li et al., 2019). Fourth, we evaluate CoCo on randomized uplift data, the official MSLR-WEB10K Fold1 split, a sequential online simulation, and Open Bandit logs. These experiments examine the offline utility–calibration trade-off, sequential gate updates, and calibration with randomized gate data.

## 2 Related Work

**Counterfactual and online learning-to-rank.** Counterfactual LTR learns and evaluates rankers from logged interactions by correcting position, exposure, and trust biases with propensity-weighted or doubly robust estimators (Joachims et al., 2017; Agarwal et al., 2019; Oosterhuis & de Rijke, 2021; Kiyohara et al., 2022). Recent work emphasizes robustness, safety, and multi-stage structure in counterfactual ranking (Niu et al., 2024; Gupta et al., 2024; 2025). Online LTR instead updates policies through interaction using dueling or multileaving feedback, differentiable click-derived objectives, contextual bandit methods, and counterfactual online learning (Yue & Joachims, 2009; Oosterhuis & de Rijke, 2018; Li et al., 2019; Zhuang & Zuccon, 2020; Zhou et al., 2023). Whole-page ranking further highlights interactions among page regions and modules (Qin et al., 2024). CoCo complements these methods by calibrating candidate eligibility before the final ranking stage.

**Conformal risk control and uplift modeling.** Conformal prediction provides distribution-free uncertainty quantification under exchangeability, and conformalized quantile regression adapts intervals to heteroscedastic errors (Angelopoulos & Bates, 2023; Romano et al., 2019). Conformal risk control and learn-then-test methods extend conformal ideas from coverage to user-specified risks (Bates et al., 2021; Angelopoulos et al., 2021; 2022). Recent work studies cross-validation, non-exchangeable, and localized variants (Cohen et al., 2024; Farinhas et al., 2024). CoCo uses a one-sided split-conformal construction for false-suppression risk. CoCo accepts uplift estimates such as T-learners and doubly robust pseudo-outcomes fit to randomized or propensity-supported data (Gutierrez & Gérardy, 2017; Künzel et al., 2019; Chen et al., 2024). Our contribution is a calibrated decision layer that converts these uplift predictions into conservative suppression rules.

## 3 Problem Formulation

### 3.1 Suppression and Ranking Decisions

**Contexts and candidates.** Let $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ denote a request context, such as a query, user state, session, page, device, or marketplace context. Let $\mathcal{W}(\boldsymbol{x}) = \{w_1, \ldots, w_m\}$ be the candidate set for context $\boldsymbol{x}$. A candidate may be an organic result, product, advertisement, widget, carousel, or recommendation module. Each candidate $w$ has an item-feature vector $\boldsymbol{v}(w) \in \mathbb{R}^{d_w}$, and the context-item feature vector is $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{v}(w)) \in \mathcal{Z}$.

**Locations, gates, and impressions.** Let $\mathcal{L} = \{\ell_1, \ldots, \ell_K\}$ be the set of nonempty display locations or slots, and let $\mathcal{L}_0 = \{\varnothing, \ell_1, \ldots, \ell_K\}$ include the virtual not-display location $\varnothing$. The location decision $L(\boldsymbol{x}, w) \in \mathcal{L}_0$ specifies where candidate $w$ is placed. The realized display indicator is $A(\boldsymbol{x}, w) = \mathbf{1}\{L(\boldsymbol{x}, w) \neq \varnothing\}$. Thus $A(\boldsymbol{x}, w) = 1$ means that the candidate item is actually shown at a nonempty slot, and $A(\boldsymbol{x}, w) = 0$ means that it is not displayed. Eligibility is controlled by a separate rule $h : \mathcal{Z} \to \{0, 1\}$ such that $h(\boldsymbol{x}, w) = 0$ means suppressing item $w$ and $h(\boldsymbol{x}, w) = 1$ means $w$ is impressed (display-eligible). Suppression precludes display, whereas impression only establishes eligibility: $h(\boldsymbol{x}, w) = 0 \Rightarrow A(\boldsymbol{x}, w) = 0$, whereas $h(\boldsymbol{x}, w) = 1$ need not imply $A(\boldsymbol{x}, w) = 1$.

**Suppression-aware policies.** A suppression-aware policy $\pi$ induces a location assignment $L_\pi(\boldsymbol{x}, w) \in \mathcal{L}_0, \forall w \in \mathcal{W}(\boldsymbol{x})$. In the score-based case used in most of this paper, we write $\pi = (h, s_{\boldsymbol{\theta}})$, where $h$ is the suppression gate and $s_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathbb{R}$ is a ranking score. If $h(\boldsymbol{x}, w) = 0$, then $L_\pi(\boldsymbol{x}, w) = \varnothing$. If $h(\boldsymbol{x}, w) = 1$, then the candidate is eligible for assignment to a nonempty location according to the ranking score and list constraints. The displayed action is

$\boldsymbol{d}_\pi(\boldsymbol{x}) = \{(w, L_\pi(\boldsymbol{x}, w)) : w \in \mathcal{W}(\boldsymbol{x}), L_\pi(\boldsymbol{x}, w) \neq \varnothing\}$. When the policy is induced by parameters $\boldsymbol{\theta}$, we write $\pi_{\boldsymbol{\theta}}$ to emphasize the ranking-score parameters.

### 3.2 Causal Effects and Logged Feedback

**Causal effect.** We use uplift as the causal effect on suppression by comparing two potential outcomes for a context-candidate pair. Let $Y_\ell(\boldsymbol{z})$ be the potential outcome at each nonempty location $\ell \in \mathcal{L}$ and $Y_\varnothing(\boldsymbol{z})$ be the outcome for not-display. In the binary setting, we further define $Y_{h=1}(\boldsymbol{z})$ as the potential outcome for impression under the serving protocol used by the system, and let $Y_{h=0}(\boldsymbol{z})$ be the potential outcome for suppression. The serving protocol specifies the candidate set, ranking model, slot constraints, and replacement or refill rule used to construct the outcome. It is held fixed within each data construction and omitted from the notation. The outcomes $Y$ are application-specific: they may be clicks, conversions, satisfaction, revenue, abandonment, composite utility, or relevance-derived utility in a benchmark. The binary uplift is

$$\Delta(\boldsymbol{z}) = \mathbb{E}\left[Y_{h=1}(\boldsymbol{z}) - Y_{h=0}(\boldsymbol{z}) \mid \boldsymbol{z}\right]. \tag{1}$$

In a binary treatment dataset, impression and receiving treatment may coincide. In LTR, impression establishes eligibility for the downstream display decision.

**Offline logged data.** The offline learner observes $\mathcal{D}_n = \{(\boldsymbol{x}_i, \mathcal{W}_i, \boldsymbol{d}_i^{\log}, H_i, A_i, L_i, Y_i, \eta_i, \omega_i)\}_{i=1}^n$, where $H_i$ is the logged gate action, $A_i$ is realized display, $L_i$ is location, $Y_i$ is the outcome, $\eta_i$ is the gate propensity, and $\omega_i$ is the ranking inclusion propensity when available. The important distinction is that $A = 0$ does not imply $H = 0$: a candidate may be eligible but rank below the displayed slots.

### 3.3 Objectives and Error Criteria

Let $\lambda(\ell) \geq 0$ be a location discount or utility weight, with $\lambda(\varnothing) = 0$. For any suppression-aware policy $\pi$, the expected ranking utility is

$$V(\pi) = \mathbb{E}\left[\sum_{w \in \mathcal{W}(\boldsymbol{x})} \sum_{\ell \in \mathcal{L}} \mathbf{1}\{L_\pi(\boldsymbol{x}, w) = \ell\} \lambda(\ell) Y_\ell(\boldsymbol{z})\right].$$

In our binary gate formulation, $h(\boldsymbol{z})$ first determines eligibility. If $h(\boldsymbol{z}) = 0$, then $L_\pi(\boldsymbol{x}, w) = \varnothing$. If $h(\boldsymbol{z}) = 1$, the ranking model may assign a nonempty location, and utility is evaluated through the actual assigned location.

False suppression (FS) means suppressing a candidate with nonnegative uplift, with risk $\mathcal{R}_{\mathrm{FS}}(\pi) = \mathbb{P}\{h(\boldsymbol{z}) = 0, \Delta(\boldsymbol{z}) \geq 0\}$. Harmful impression (HI) means making a negative-uplift candidate eligible, with risk $\mathcal{R}_{\mathrm{HI}}(\pi) = \mathbb{P}\{h(\boldsymbol{z}) = 1, \Delta(\boldsymbol{z}) < 0\}$. Eligibility still need not lead to display. The two errors have different roles: CoCo controls FS for its calibration target, while utility reflects the cost of both types of error.

For an online policy $\pi_t$ and oracle $\pi_t^\star$, define ranking regret $\mathrm{Reg}_T = \sum_{t=1}^T \{V_t(\pi_t^\star) - V_t(\pi_t)\}$. Let $e_{\mathrm{FS},t}(w) = \mathbf{1}\{h_t(\boldsymbol{x}_t, w) = 0, \Delta_t(\boldsymbol{x}_t, w) \geq 0\}$ and $e_{\mathrm{HI},t}(w) = \mathbf{1}\{h_t(\boldsymbol{x}_t, w) = 1, \Delta_t(\boldsymbol{x}_t, w) < 0\}$. For $j \in \{\mathrm{FS}, \mathrm{HI}\}$, the candidate-averaged empirical error is $\hat{\mathcal{R}}_{j,T} = T^{-1} \sum_t |\mathcal{W}_t|^{-1} \sum_{w \in \mathcal{W}_t} e_{j,t}(w)$. The objective is to achieve high utility while controlling suppression risk. Target-level calibration alone does not control both errors.

## 4 Conformal Confidence Suppression for Learning-to-Rank

### 4.1 Conformal Confidence (CoCo) Suppression

**Point and conformal gates.** A point gate $h^{\mathrm{point}}(\boldsymbol{z}) = \mathbf{1}\{\widehat{\Delta}(\boldsymbol{z}) \geq 0\}$ suppresses negative uplift estimates, including beneficial candidates underestimated because of noise. CoCo uses an upper bound $U_\alpha$ and retains a candidate whenever $U_\alpha(\boldsymbol{z}) \geq 0$. The gate controls eligibility, and the ranker determines actual display.

**T-learner.** Fit outcome models $\widehat{\mu}_{1,\widehat{\boldsymbol{\beta}}_1}$ and $\widehat{\mu}_{0,\widehat{\boldsymbol{\beta}}_0}$ on factual rows with $H = 1$ and $H = 0$, respectively, using squared loss for continuous outcomes or log loss for binary outcomes. Their difference

is $\widehat{\Delta}_{\mathrm{T},\widehat{\boldsymbol{\beta}}}(\boldsymbol{z}) = \widehat{\mu}_{1,\widehat{\boldsymbol{\beta}}_1}(\boldsymbol{z}) - \widehat{\mu}_{0,\widehat{\boldsymbol{\beta}}_0}(\boldsymbol{z})$, where $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_0)$. Suppression examples must reflect an intentional gate intervention. Ordinary non-display from ranking below the slots gives $A = 0$, not necessarily $H = 0$.

**DR learner.** On a nuisance split or cross-fitting folds, fit outcome models $\widehat{\mu}_{h,\widehat{\boldsymbol{\gamma}}_h}(\boldsymbol{z}) \approx \mathbb{E}[Y \mid H = h, \boldsymbol{z}]$ and a propensity model $\widehat{\eta}_{\widehat{\boldsymbol{\rho}}}(\boldsymbol{z}) \approx \mathbb{P}(H = 1 \mid \boldsymbol{z})$. Use the known propensity when randomized logging supplies it. Out-of-fold predictions define

$$\Phi_i = \widehat{\mu}_{1,\widehat{\boldsymbol{\gamma}}_1}(\boldsymbol{z}_i) - \widehat{\mu}_{0,\widehat{\boldsymbol{\gamma}}_0}(\boldsymbol{z}_i) + \frac{H_i\{Y_i - \widehat{\mu}_{1,\widehat{\boldsymbol{\gamma}}_1}(\boldsymbol{z}_i)\}}{\widehat{\eta}_{\widehat{\boldsymbol{\rho}}}(\boldsymbol{z}_i)} - \frac{(1 - H_i)\{Y_i - \widehat{\mu}_{0,\widehat{\boldsymbol{\gamma}}_0}(\boldsymbol{z}_i)\}}{1 - \widehat{\eta}_{\widehat{\boldsymbol{\rho}}}(\boldsymbol{z}_i)}. \tag{2}$$

Regressing $\Phi_i$ on $\boldsymbol{z}_i$ gives the final uplift model $\widehat{\Delta}_{\mathrm{DR},\widehat{\boldsymbol{\beta}}}$ using targets constructed by the nuisance models. Propensity clipping may stabilize implementation, but exact double robustness requires the conditions in Proposition 1.

**Calibration target.** CoCo calibrates an observable or constructed target $\Gamma_i$, rather than the generally unobserved conditional mean $\Delta(\boldsymbol{z}_i)$. A justified matched design uses $\Gamma_i^{\mathrm{match}} = Y_i^{\mathrm{imp}} - Y_i^{\mathrm{sup}}$ from comparable impressed and suppressed opportunities. DR calibration uses $\Gamma_i^{\mathrm{DR}} = \Phi_i$, while a full-information benchmark may supply utility directly. Matching is not needed to fit a T-learner, but its calibration target still needs justification. Unbiasedness $\mathbb{E}[\Gamma \mid \boldsymbol{z}] = \Delta(\boldsymbol{z})$ identifies the target's conditional mean, but it does not make coverage of noisy $\Gamma$ a confidence guarantee for $\Delta$.

**One-sided calibration.** Using $m$ calibration pairs excluded from fitting the calibrated functions, form $S_i = \widehat{\Delta}(\boldsymbol{z}_i) - \Gamma_i$. With $k_\alpha = \lfloor (m+1)\alpha \rfloor$, set $q_\alpha = S_{(k_\alpha)}$, taking $q_\alpha = -\infty$ when $k_\alpha = 0$. The gate and upper prediction bound are

$$U_\alpha(\boldsymbol{z}) = \widehat{\Delta}(\boldsymbol{z}) - q_\alpha, \qquad h_\alpha^{\mathrm{coco}}(\boldsymbol{z}) = \mathbf{1}\{U_\alpha(\boldsymbol{z}) \geq 0\}. \tag{3}$$

Since $S_{\mathrm{new}} \geq q_\alpha$ is equivalent to $\Gamma_{\mathrm{new}} \leq U_\alpha(\boldsymbol{z}_{\mathrm{new}})$, exchangeability gives upper coverage at least $1-\alpha$. False suppression of a nonnegative target requires this coverage event to fail. Thus calibration controls target-level false suppression, while utility remains an empirical objective.

**CoCo variants and display formulations.** A CoCo implementation combines an uplift estimator (T or DR), a calibration variant, and a display granularity (context-item or candidate-slot). The three variants differ in how they adapt the calibration correction.

**CoCo-1: global calibration.** CoCo-1 pools all calibration scores and computes a single lower-tail quantile $q_\alpha$. Every candidate receives the same additive correction:

$$U_\alpha^1(\boldsymbol{z}) = \widehat{\Delta}(\boldsymbol{z}) - q_\alpha.$$

Global calibration is suitable for prediction errors with similar distributions across candidates. Coverage is marginal over the calibration population and need not hold within every subgroup.

**CoCo-M: multi-group calibration.** CoCo-M assigns a group $G(\boldsymbol{z})$, such as query type, user segment, item type, or a score bin fixed before calibration. For each group $g$, it computes $q_{\alpha,g}$ from scores with $G(\boldsymbol{z}_i) = g$ and uses

$$U_\alpha^{\mathrm{M}}(\boldsymbol{z}) = \widehat{\Delta}(\boldsymbol{z}) - q_{\alpha,G(\boldsymbol{z})}.$$

This adapts to systematic differences between groups and gives group-conditional validity under within-group exchangeability. Sparse groups may be merged or assigned a prespecified global fallback. The fallback does not generally preserve group-conditional coverage.

**CoCo-F: locally adaptive calibration.** CoCo-F fits a positive scale $\widehat{\sigma}(\boldsymbol{z})$ for local prediction-error magnitude using non-calibration data. It calibrates normalized scores $R_i = S_i/\widehat{\sigma}(\boldsymbol{z}_i)$, obtains a global quantile $q_{\alpha,\mathrm{norm}}$, and forms

$$U_\alpha^{\mathrm{F}}(\boldsymbol{z}) = \widehat{\Delta}(\boldsymbol{z}) - q_{\alpha,\mathrm{norm}}\widehat{\sigma}(\boldsymbol{z}).$$

When the lower-tail quantile is negative, higher estimated uncertainty produces a wider upper bound. This is useful for heteroscedastic errors: the correction varies continuously within as well as between groups. Validity requires exchangeable normalized scores and a scale function fitted independently of calibration. All three variants use the gate in Eq. (3).

### 4.2 OFFLINE COCO-LTR

A fixed CoCo gate and ranking score $f_{\boldsymbol{\theta}}$ define $\pi_{\boldsymbol{\theta}} = (h_\alpha^{\mathrm{coco}}, f_{\boldsymbol{\theta}})$. For request $\boldsymbol{x}$, rank $\mathcal{E}_\alpha(\boldsymbol{x}) = \{w \in \mathcal{W}(\boldsymbol{x}) : U_\alpha(\boldsymbol{x}, w) \geq 0\}$ by $f_{\boldsymbol{\theta}}$, assigning at most $K$ candidates to nonempty locations. If fewer candidates remain eligible, leave the remaining slots empty. This gate-before-ranking order separates eligibility from display.

**Offline ranker-learning objective.** For a policy $\pi_{\boldsymbol{\theta}}$, define the induced location weight $\kappa_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{x}, w) = \mathbf{1}\{L_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{x}, w) \neq \varnothing\}\lambda(L_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{x}, w))$. Let $\widehat{\omega}_i(w)$ be the probability that candidate $w$ is included in the logged displayed action for request $i$. An inverse propensity score (IPS) estimator for the gated ranking value is

$$\widehat{V}_{\mathrm{IPS}}(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{w\in\mathcal{W}_i}\frac{A_i(w)\kappa_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{x}_i, w)Y_i(w)}{\widehat{\omega}_i(w)}.$$

A doubly robust value estimator is

$$\widehat{V}_{\mathrm{DR}}(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{w\in\mathcal{W}_i}\kappa_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{x}_i, w)\left[\widehat{m}_1(\boldsymbol{x}_i, w) + \frac{A_i(w)}{\widehat{\omega}_i(w)}\{Y_i(w) - \widehat{m}_1(\boldsymbol{x}_i, w)\}\right]. \tag{4}$$

The regression $\widehat{m}_1$ in Eq. (4) is a displayed-outcome model for ranking-policy value estimation. It is separate from the T-learner or DR-learner nuisance functions used to estimate impression uplift, even if an implementation reuses features or model classes.

After fixing the gate construction, learn or select $\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}}\{\widehat{V}_{\mathrm{DR}}(\pi_{\boldsymbol{\theta}}) - \rho_\Omega\Omega(\boldsymbol{\theta})\}$. Here $\rho_\Omega \geq 0$ controls regularization, with $\Omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2/2$ corresponding to ridge or weight decay. Replacing DR by IPS gives a propensity-only objective. The MSLR experiment evaluates the gate with a fixed ranker and does not test optimization of this logged-data objective.

### 4.3 ONLINE COCO-LTR

Online CoCo-LTR maintains a ranker state $\psi_t$, fitted T/DR and scale models $\mathcal{M}_t$, and a buffer $\mathcal{B}_t$ of pre-update records $(S_i, R_i, g_i)$. Let $Q_\alpha(\mathcal{S})$ denote the lower-tail order statistic defined in Section 4.1, with the same zero-index convention. Algorithm 1 computes the gate before feedback, ranks at most $K$ eligible candidates using the base ranker's policy $\mathcal{A}_{\mathrm{rank}}$, and then updates the buffer and models.

**Feedback and updates.** Feedback $\mathcal{O}_t$ contains available outcomes, actions $H_t, A_t, L_t$, and actual logging propensities $\eta_t, \omega_t$. Let $\mathcal{J}_t$ index valid target observations: $\mathcal{J}_t = \mathcal{W}_t$ is possible under full information, whereas bandit targets require a justified T/DR construction. Gate exploration uses eligibility probability $p_t(\boldsymbol{z}) \in [p_{\min}, 1 - p_{\min}]$ in a bucket of probability $\varepsilon_t$. The nominal gate uses $\varepsilon_t = 0$. Ranking exploration alone does not identify suppression effects.

The update operators follow the chosen ranker and T/DR training rules, with scale fitting based on pre-update residual magnitudes. Cached scores avoid fitting on the current feedback before evaluation, but evolving predictors need not yield exchangeable scores. Freezing fitted functions while collecting an independent calibration window is a split-calibrated alternative. Section 5.3 states the assumptions needed for validity under rolling updates (Barber et al., 2023). Appendix C.7 specifies the evaluated full-information implementation, preprocessing, and quantile convention.

**Algorithm 1** Online CoCo-LTR with rolling calibration

**Require:** $v \in \{1, \mathrm{M}, \mathrm{F}\}$, $\alpha$, $m$, $K$, warm-start data $\mathcal{D}_0$, $\{\varepsilon_t\}_{t=1}^{T}$.

$(\psi_1, \mathcal{M}_1, \mathcal{B}_1) \leftarrow \mathrm{WarmStart}(\mathcal{D}_0)$
**for** $t = 1, \ldots, T$ **do**
  **for** $w \in \mathcal{W}_t$ **do**
    $\boldsymbol{z}_{tw} \leftarrow (\boldsymbol{x}_t, w), \quad d_{tw} \leftarrow \widehat{\Delta}_t(\boldsymbol{z}_{tw}), \quad g_{tw} \leftarrow G(\boldsymbol{z}_{tw})$
    $(\mathcal{S}_{tw}, a_{tw}) \leftarrow \begin{cases} (\{S_i : i \in \mathcal{B}_t\}, 1), & v = 1, \\ (\{S_i : i \in \mathcal{B}_t,\ g_i = g_{tw}\}, 1), & v = \mathrm{M}, \\ (\{R_i : i \in \mathcal{B}_t\}, \widehat{\sigma}_t(\boldsymbol{z}_{tw})), & v = \mathrm{F} \end{cases}$
    $U_{tw} \leftarrow d_{tw} - a_{tw} Q_\alpha(\mathcal{S}_{tw}), \quad h_{tw}^{\mathrm{coco}} \leftarrow \mathbf{1}\{U_{tw} \geq 0\}$
  **end for**
  $\xi_t \sim \mathrm{Bernoulli}(\varepsilon_t)$
  **if** $\xi_t = 0$ **then** $H_{tw} \leftarrow h_{tw}^{\mathrm{coco}}$; **else** $H_{tw} \sim \mathrm{Bernoulli}(p_t(\boldsymbol{z}_{tw})),\ \forall w$
  $\eta_{tw} \leftarrow (1 - \varepsilon_t) h_{tw}^{\mathrm{coco}} + \varepsilon_t p_t(\boldsymbol{z}_{tw}), \quad w \in \mathcal{W}_t$
  $\mathcal{E}_t \leftarrow \{w \in \mathcal{W}_t : H_{tw} = 1\}, \quad (L_t, \omega_t) \sim \mathcal{A}_{\mathrm{rank}}(\psi_t, \mathcal{E}_t, K)$
  $\mathcal{O}_t \leftarrow \mathrm{Feedback}(H_t, L_t), \quad \mathcal{J}_t \leftarrow \mathrm{ValidTargets}(\mathcal{O}_t)$
  **for** $w \in \mathcal{J}_t$ **do**
    $\Gamma_{tw} \leftarrow \mathrm{Target}_{\mathrm{T/DR}}(\mathcal{M}_t, \mathcal{O}_t; \boldsymbol{z}_{tw})$
    $S_{tw} \leftarrow d_{tw} - \Gamma_{tw}, \quad R_{tw} \leftarrow S_{tw} / \widehat{\sigma}_t(\boldsymbol{z}_{tw})$
  **end for**
  $\mathcal{B}_{t+1} \leftarrow \mathrm{Tail}_m\big(\mathcal{B}_t \,\|\, [(S_{tw}, R_{tw}, g_{tw})]_{w \in \mathcal{J}_t}\big)$
  $\psi_{t+1} \leftarrow \mathrm{Update}_{\mathrm{rank}}(\psi_t, \mathcal{O}_t)$
  $\mathcal{M}_{t+1} \leftarrow \mathrm{Update}_{\mathrm{T/DR,scale}}(\mathcal{M}_t, \mathcal{O}_t)$
**end for**

$\mathrm{Tail}_m$ retains the latest $m$ records. CoCo-M uses its prespecified sparse-group fallback. For $v \neq \mathrm{F}$, normalized scores can be omitted.

# 5 Theoretical Analysis

## 5.1 Split-Conformal Guarantees for CoCo Suppression

**Assumption 1** (Split calibration)**.** *The data used to fit $\widehat{\Delta}$ and any nuisance or scale functions are independent of the calibration examples used to compute $q_\alpha$.*

A held-out calibration split prevents overfitting the scores used to quantify uncertainty.

**Assumption 2** (Exchangeable calibration target)**.** *Conditional on the fitted functions, the calibration pairs $\{(\boldsymbol{z}_i, \Gamma_i)\}_{i \in \mathcal{D}_{\mathrm{cal}}}$ and a future pair $(\boldsymbol{z}_{\mathrm{new}}, \Gamma_{\mathrm{new}})$ are exchangeable.*

Exchangeability does not require a correctly specified uplift model, but distribution shift or adaptive logging can invalidate it. The guarantee concerns $\Gamma$. For a DR pseudo-outcome, causal interpretation additionally requires the gate-level assumptions and data construction below.

**Theorem 1** (Generic CoCo false-suppression control)**.** *Let $S_i = \widehat{\Delta}(\boldsymbol{z}_i) - \Gamma_i$ be the one-sided nonconformity scores. Let $q_\alpha$ be the lower-tail split-conformal $\alpha$-quantile in Eq.* (3)*. Define $U_\alpha(\boldsymbol{z}) = \widehat{\Delta}(\boldsymbol{z}) - q_\alpha$ and $h_\alpha^{\mathrm{coco}}(\boldsymbol{z}) = \mathbf{1}\{U_\alpha(\boldsymbol{z}) \geq 0\}$. Under Assumptions 1 and 2,*

$$\mathbb{P}\{\Gamma_{\mathrm{new}} \leq U_\alpha(\boldsymbol{z}_{\mathrm{new}})\} \geq 1 - \alpha.$$

*Moreover,*

$$\mathbb{P}\{h_\alpha^{\mathrm{coco}}(\boldsymbol{z}_{\mathrm{new}}) = 0,\ \Gamma_{\mathrm{new}} \geq 0\} \leq \alpha.$$

**Corollary 1** (T-CoCo with matched uplift targets)**.** *If $\Gamma_i = \Gamma_i^{\mathrm{match}} = Y_i^{\mathrm{imp}} - Y_i^{\mathrm{sup}}$ is a matched noisy uplift target and the matched calibration scores are exchangeable with the future matched score, then Theorem 1 gives false-suppression control for T-CoCo using $\widehat{\Delta}_{\mathrm{T}}$.*

**Assumption 3** (Consistency, overlap, and ignorability)**.** *For the binary gate construction, the observed outcome equals the potential outcome under the realized gate action $H$. There exists $\epsilon > 0$ such that $\epsilon \leq \eta(\boldsymbol{z}) \leq 1 - \epsilon$, and $(Y_{h=1}, Y_{h=0}) \perp H \mid \boldsymbol{z}$ under the serving protocol used by the data construction.*

**Proposition 1** (Double robustness of the DR uplift target)**.** *Under Assumption 3, if either $\widehat{\eta}_{\widehat{\boldsymbol{\rho}}} = \eta$ or $(\widehat{\mu}_1, \widehat{\mu}_0) = (\mu_1, \mu_0)$, then the pseudo-outcome in Eq.* (2) *satisfies $\mathbb{E}[\Phi \mid \boldsymbol{z}] = \Delta(\boldsymbol{z})$.*

**Corollary 2** (DR-CoCo pseudo-uplift control)**.** *If $\Gamma_i = \Phi_i$ and the DR calibration scores $\widehat{\Delta}_{\mathrm{DR}}(\boldsymbol{z}_i) - \Phi_i$ are exchangeable with the future score, then Theorem 1 gives pseudo-uplift false-suppression control for DR-CoCo. Under Proposition 1, this pseudo-uplift target has conditional mean equal to the causal impression/suppression uplift.*

**Proposition 2** (CoCo-M and CoCo-F validity)**.** *For CoCo-M, if calibration and test scores are exchangeable conditional on group $G = g$, then Theorem 1 holds conditional on $G = g$. For CoCo-F, if the normalized scores $S_i/\widehat{\sigma}(\boldsymbol{z}_i)$ and the future normalized score are exchangeable conditional on the fitted scale function, then Theorem 1 holds with $q_\alpha(\boldsymbol{z}) = q_{\alpha,\mathrm{norm}}\widehat{\sigma}(\boldsymbol{z})$.*

### 5.2 Offline Gated-Ranking Guarantees

Once the selected CoCo gate is fixed relative to the evaluation sample, the gated ranking policy can be analyzed using off-policy evaluation. The selected gate is represented by $h_\alpha^{\mathrm{coco}}$.

**Theorem 2** (List-level false-suppression bound)**.** *If a request contains at most $m_{\max}$ calibrated candidate gates, each with target false-suppression probability at most $\alpha_{\mathrm{item}}$, the probability of any such error is at most $m_{\max}\alpha_{\mathrm{item}}$. Set $\alpha_{\mathrm{item}} = \alpha_{\mathrm{list}}/m_{\max}$ for list-level control. The candidate count, not the number of display slots $K$, determines this union bound.*

**Proposition 3** (Double robustness of gated off-policy ranking value)**.** *Suppose $h_\alpha^{\mathrm{coco}}$ and $\pi_{\boldsymbol{\theta}}$ are fixed with respect to the evaluation sample. Under consistency and overlap for displayed outcomes, the estimator in Eq.* (4) *is unbiased for the gated ranking value $V(\pi_{\boldsymbol{\theta}})$ if either the logging inclusion model $\widehat{\omega}$ is correct or the displayed-outcome model $\widehat{m}_1$ is correct.*

**Theorem 3** (Finite-class offline policy selection bound)**.** *Let $\Pi$ be a finite class of gated ranking policies, rewards and location weights lie in $[0, 1]$, at most $K$ items be displayed, and logging inclusion probabilities satisfy $\omega(w) \geq \epsilon_\omega > 0$ for contributing items. Define*

$$\varepsilon_n = \frac{K}{\epsilon_\omega}\sqrt{\frac{\log(2|\Pi|/\delta)}{2n}}.$$

*With probability at least $1 - \delta$, $\sup_{\pi\in\Pi} |\widehat{V}_{\mathrm{IPS}}(\pi) - V(\pi)| \leq \varepsilon_n$ and $V(\pi^\star) - V(\widehat{\pi}) \leq 2\varepsilon_n$, where $\widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{V}_{\mathrm{IPS}}(\pi)$ and $\pi^\star \in \arg\max_{\pi\in\Pi} V(\pi)$.*

### 5.3 Rolling Calibration and Online Regret

**Theorem 4** (One-step and average target-risk control)**.** *At round $t$, suppose the $m_t$ active scores and current score are exchangeable conditional on fitted objects and prespecified design information $\mathcal{F}_t$. The information $\mathcal{F}_t$ must not fix all realized calibration scores. If the threshold is their calibration order statistic $S_{(k_t)}$, with $0 \leq k_t \leq m_t$ and the same $k_t = 0$ convention, then, for a calibrated candidate,*

$$\mathbb{P}\{h_t^{\mathrm{coco}} = 0, \Gamma_t \geq 0 \mid \mathcal{F}_t\} \leq \frac{k_t}{m_t + 1} =: a_t.$$

*Consequently $\mathbb{E}[T^{-1}\sum_{t=1}^T \mathbf{1}\{h_t^{\mathrm{coco}} = 0, \Gamma_t \geq 0\}] \leq T^{-1}\sum_t \mathbb{E}[a_t]$. The floor convention gives $a_t \leq \alpha$, whereas a ceiling adds less than $1/(m_t+1)$. A gate-exploration probability $\varepsilon_t$ adds at most $\varepsilon_t$ to the served-policy bound.*

This is a marginal-in-time expectation statement, not a guarantee conditional on the entire observed history or a simultaneous guarantee at every round. It applies to CoCo-M within valid groups and to CoCo-F under exchangeability of normalized scores. Model drift, cached scores, and shared fitting/calibration data can violate these assumptions. The online simulation evaluates performance under sequential updates empirically.

**Proposition 4** (Suppression-aware regret decomposition)**.** *Let $\pi_t^{\mathrm{gate}\star}$ maximize value over the gated feasible set and let $\mathrm{Reg}_T^{\mathrm{base}} = \sum_t\{V_t(\pi_t^{\mathrm{gate}\star}) - V_t(\pi_t)\}$. Assume the loss of feasible value caused by the gate is at most $B(M_{\mathrm{FS},t} + M_{\mathrm{HI},t})$, where these are the candidate-averaged error fractions defining $\hat{\mathcal{R}}_{\mathrm{FS},T}, \hat{\mathcal{R}}_{\mathrm{HI},T}$. Then*

$$\mathrm{Reg}_T \leq \mathrm{Reg}_T^{\mathrm{base}} + BT(\hat{\mathcal{R}}_{\mathrm{FS},T} + \hat{\mathcal{R}}_{\mathrm{HI},T}).$$

The value-loss assumption is additional to calibration: $B$ must absorb candidate count and interaction effects. Theorem 4 controls errors defined by $\Gamma$, whereas the regret fractions use $\Delta$. Relating these targets and controlling harmful impressions require further assumptions. Thus fixed-$\alpha$ calibration alone does not imply sublinear ranking regret.

## 6 Experiments

We evaluate utility and ranking quality on MSLR, sequential updates in an online simulation, and target calibration on randomized Criteo data. Ungated rankers and matched point gates separate suppression gains from calibration effects. Open Bandit provides supporting OPE evidence (Appendix C.6). Main comparisons use $\alpha = 0.10$, with sensitivity on $\{0.05, 0.10, 0.20\}$. We report mean $\pm$ standard error (SE) across seeds.

### 6.1 Offline Utility and Calibration on MSLR

We use official MSLR-WEB10K Fold1 splits, a fixed net-utility ranker, and display cost $0.10$. This full-information experiment evaluates gating, not causal identification or DR ranker training. Table 1 reports net Utility@10 and relevance-based NDCG@10, using all candidates for the ideal ranking. DR+CoCo-M improves utility over the ungated ranker by $11.6\%$ ($+0.0295 \pm 0.0009$) and retains $93.1\%$ of its NDCG. Point has higher utility but retains $71.2\%$ of the ranker's NDCG.

Table 1: MSLR, three seeds. Utility includes display cost; NDCG measures relevance without that cost. Violation is $\Pr_n(\Gamma > U_\alpha)$, in percent. Bold marks the highest utility or NDCG among all listed rules within each learner.

| | DR gate | | | T gate | | |
|---|---|---|---|---|---|---|
| Rule | Utility@10 | NDCG@10 | Viol. | Utility@10 | NDCG@10 | Viol. |
| Ranker | $0.2553 \pm 0.0001$ | $\mathbf{0.4487} \pm 0.0001$ | – | $0.2553 \pm 0.0001$ | $\mathbf{0.4487} \pm 0.0001$ | – |
| Point | $\mathbf{0.3104} \pm 0.0013$ | $0.3193 \pm 0.0008$ | – | $\mathbf{0.3128} \pm 0.0007$ | $0.3230 \pm 0.0005$ | – |
| CoCo-1 | $0.2553 \pm 0.0001$ | $\mathbf{0.4487} \pm 0.0001$ | 9.8 | $0.2553 \pm 0.0001$ | $\mathbf{0.4487} \pm 0.0001$ | 9.8 |
| CoCo-M | $0.2848 \pm 0.0010$ | $0.4179 \pm 0.0008$ | 9.8 | $0.2846 \pm 0.0009$ | $0.4174 \pm 0.0004$ | 9.9 |
| CoCo-F | $0.2708 \pm 0.0005$ | $0.4363 \pm 0.0002$ | 9.8 | $0.2707 \pm 0.0002$ | $0.4360 \pm 0.0006$ | 9.7 |

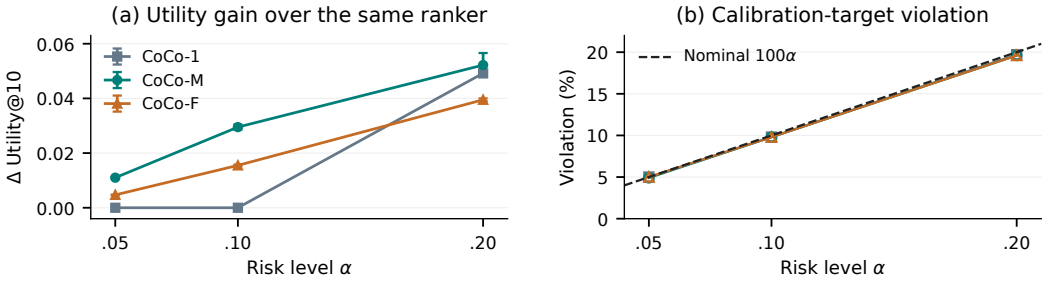


Figure 1: MSLR net ranker with DR gates at cost $0.10$. Gains ($\pm 1$ SE) increase as the risk budget grows, while target violation tracks $\alpha$. CoCo-M improves on global CoCo-1 at the conservative budgets, with nearly overlapping calibration curves.

DR+CoCo-M exceeds Point's NDCG by $0.0986 \pm 0.0004$ across paired seeds, with utility $0.2848$ versus $0.3104$. CoCo-F preserves more relevance with smaller utility gains over the ranker. T gates show the same ordering. These results support a utility–ranking-quality trade-off. Figure 1 shows near-nominal target violation. Point's false-suppression risk was not recorded, precluding a matched-risk comparison. Appendix C.5 reports suppression, both rankers, and alpha and cost sensitivity.

### 6.2 Sequential Online CoCo-LTR

The synthetic stream has 900 queries, 24 candidates/query, $K = 10$, three seeds, and a distribution change at query index 540. Evaluation uses the last 660 queries. A shared linear ranker, T/DR learners, scale models, and cached calibration scores update after each query. The buffer holds 120 queries. All methods use the same ranker state and stream to isolate the effect of the gate. Features include the regime indicator and are standardized using the full stream, so this is a controlled sequential simulation with access to future feature statistics.

Table 2: Online simulation, three seeds. Exceedance is $\Pr_n(\Delta > U_\alpha)$ against latent utility, not calibration-target violation. Suppression is the fraction of ungated top-10 candidates removed. Rates are percentages. Bold indicates the highest utility among conformal variants.

| | DR gate | | | T gate | | |
|---|---|---|---|---|---|---|
| Rule | Utility/query | Exceed. | Supp. | Utility/query | Exceed. | Supp. |
| Ranker | $0.182 \pm 0.013$ | – | 0.0 | $0.182 \pm 0.013$ | – | 0.0 |
| Point | $0.632 \pm 0.016$ | – | 47.1 | $0.639 \pm 0.013$ | – | 46.0 |
| CoCo-1 | $0.226 \pm 0.016$ | 0.87 | 2.0 | $0.263 \pm 0.012$ | 2.45 | 3.6 |
| CoCo-M | $\mathbf{0.394} \pm 0.008$ | 2.46 | 9.7 | $\mathbf{0.393} \pm 0.011$ | 3.44 | 9.8 |
| CoCo-F | $0.323 \pm 0.010$ | 1.36 | 6.3 | $0.365 \pm 0.011$ | 3.61 | 8.7 |

DR+CoCo-M gains $0.2120 \pm 0.0053$ utility/query over the ranker, and T+CoCo-M gains $0.2110 \pm 0.0079$. Figure 2a compares DR utility and suppression. Point achieves higher utility and suppresses $47.1\%$ of the ungated top-10, versus $9.7\%$ for CoCo-M. Lower suppression alone does not establish lower false-suppression risk. These results show gains over the ungated ranker under sequential updates, without establishing an advantage over Point. This full-information simulation does not establish noisy-target coverage, a regret rate, or performance with bandit feedback. Appendix C.7 reports the cumulative curves (Figure 5) and implementation details.

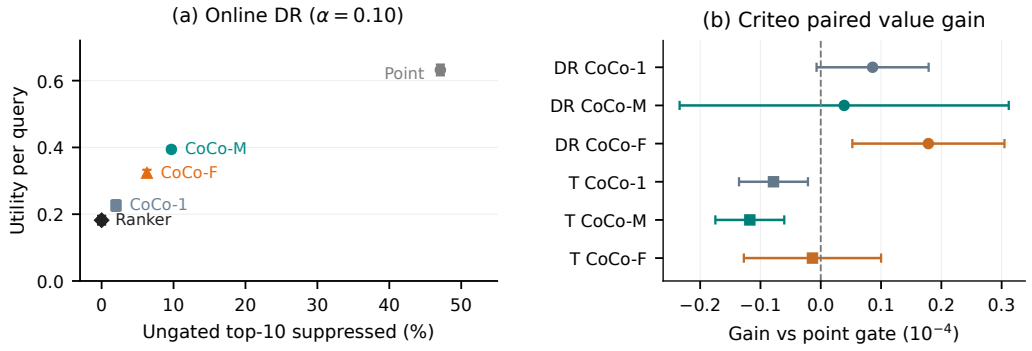


Figure 2: (a) Online DR utility versus suppression of the ungated top-10. Vertical bars show $\pm 1$ SE across three seeds. Conformal gates use $\alpha = 0.10$. (b) Criteo gains over matched point gates, with $\pm 1$ SE across five seeds.

### 6.3 Randomized Gate Evidence on Criteo

Criteo Uplift v2.1 supplies a randomized binary gate-arm analogue. DR+CoCo-F improves mean value by $2.04\%$, with a paired gain of $(0.179 \pm 0.126) \times 10^{-4}$. T variants do not improve on the point gate (Figure 2b). Empirical target violation is approximately $9.5$–$10.1\%$. The randomized data thus show near-nominal target calibration, although evidence of a value gain is limited. Appendix C.4 reports all variants and sensitivity to the risk level.

## AI USE STATEMENT

Generative AI tools were used in this work to support research-methodology and experiment-design feedback, implementation and debugging of research code, and proof development, interpretation of experimental results, literature search. All AI-assisted code, derivations, citations, experiments, and manuscript text were reviewed and validated by the authors; reported experimental results were produced by the authors' research pipeline rather than generated by the AI tools. The authors take responsibility for the final content of this work.

## REFERENCES


Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. A general framework for counterfactual learning-to-rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19, pp. 5–14. ACM, 2019. doi: 10.1145/3331184.3331202. URL https://arxiv.org/abs/1805.00065.

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. *arXiv preprint arXiv:2110.01052*, 2021. URL https://arxiv.org/abs/2110.01052.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022. URL https://arxiv.org/abs/2208.02814.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2023. URL https://arxiv.org/abs/2202.13415.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 2021. URL https://arxiv.org/abs/2101.02703. Also available as arXiv:2101.02703.

Jiaju Chen, Wang Wenjie, Chongming Gao, Peng Wu, Jianxiong Wei, and Qingsong Hua. Treatment effect estimation for user interest exploration on recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1861–1871, 2024.

Kfir M. Cohen, Sangwoo Park, Osvaldo Simeone, and Shlomo Shamai. Cross-validation conformal risk control. *arXiv preprint arXiv:2401.11974*, 2024. URL https://arxiv.org/abs/2401.11974.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VnAJHnYniV.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *arXiv preprint arXiv:2106.00170*, 2021. URL https://arxiv.org/abs/2106.00170.

Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. Practical and robust safety guarantees for advanced counterfactual learning to rank. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 737–747. ACM, 2024. doi: 10.1145/3627673.3679531. URL https://arxiv.org/abs/2407.19943.

Shashank Gupta, Yiming Liao, and Maarten de Rijke. Towards two-stage counterfactual learning to rank. *arXiv preprint arXiv:2506.20854*, 2025. URL https://arxiv.org/abs/2506.20854.

Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*, pp. 1–13. PMLR, 2017.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pp. 781–789. ACM, 2017. URL https://arxiv.org/abs/1608.04468.

Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, Nobuyuki Shimizu, and Yasuo Yamamoto. Doubly robust off-policy evaluation for ranking policies under the cascade behavior model. *arXiv preprint arXiv:2202.01562*, 2022. URL https://arxiv.org/abs/2202.01562.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. URL https://arxiv.org/abs/1706.03461.

Hang Li. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862, 2011.

Shuai Li, Tor Lattimore, and Csaba Szepesvári. Online learning to rank with features. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3856–3865. PMLR, 2019. URL https://arxiv.org/abs/1810.02567.

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

Zechun Niu, Jiaxin Mao, Qingyao Ai, and Ji-Rong Wen. Investigating the robustness of counterfactual learning to rank models: A reproducibility study. *arXiv preprint arXiv:2404.03707*, 2024. URL https://arxiv.org/abs/2404.03707.

Harrie Oosterhuis and Maarten de Rijke. Differentiable unbiased online learning to rank. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1293–1302, 2018.

Harrie Oosterhuis and Maarten de Rijke. Unifying online and counterfactual learning to rank: A novel counterfactual estimator that effectively utilizes online interventions. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pp. 463–471. ACM, 2021. doi: 10.1145/3437963.3441794. URL https://arxiv.org/abs/2012.04426.

Zhou Qin, Kai Yuan, Pratik Lahiri, and Wenyang Liu. Cooperative multi-agent deep reinforcement learning in content ranking optimization. *arXiv preprint arXiv:2408.04251*, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.03222.

Daria Sorokina and Erick Cantu-Paz. Amazon search: The joy of ranking products. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 459–460, 2016.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208, 2009.

Tianchen Zhou, Jia Liu, Yang Jiao, Chaosheng Dong, Yetian Chen, Yan Gao, and Yi Sun. Bandit learning to rank with position-based click models: Personalized and equal treatments. *arXiv preprint arXiv:2311.04528*, 2023. URL https://arxiv.org/abs/2311.04528.

Shengyao Zhuang and Guido Zuccon. Counterfactual online learning to rank. In *Advances in Information Retrieval*, pp. 415–430. Springer, 2020. doi: 10.1007/978-3-030-45439-5_28. URL https://doi.org/10.1007/978-3-030-45439-5_28.

# A PROOFS

## A.1 PROOF OF THEOREM 1

Let the calibration set contain $m$ examples and write

$$S_i = \widehat{\Delta}(\boldsymbol{z}_i) - \Gamma_i, \qquad i = 1, \dots, m,$$

for the calibration scores. Let

$$S_{m+1} = S_{\text{new}} = \widehat{\Delta}(\boldsymbol{z}_{\text{new}}) - \Gamma_{\text{new}}$$

be the future score. By Assumptions 1–2, conditional on the fitted functions, the random variables

$$S_1, \dots, S_m, S_{m+1}$$

are exchangeable.

Let $S_{(1)} \le \cdots \le S_{(m)}$ be the ordered calibration scores. For the lower-tail split-conformal quantile, take

$$k_\alpha = \lfloor (m+1)\alpha \rfloor, \qquad q_\alpha = \begin{cases} S_{(k_\alpha)}, & k_\alpha \ge 1, \\ -\infty, & k_\alpha = 0. \end{cases}$$

If $k_\alpha = 0$, then $q_\alpha = -\infty$, $U_\alpha(\boldsymbol{z}) = +\infty$, and the coverage statement is immediate. We therefore consider $k_\alpha \ge 1$.

Let $R$ be the rank of $S_{m+1}$ among the $m+1$ exchangeable scores, using a fixed deterministic tie-breaking rule. Exchangeability implies

$$\mathbb{P}\{R = r \mid \widehat{\Delta}, \widehat{\sigma}, \text{nuisance fits}\} = \frac{1}{m+1}, \qquad r = 1, \dots, m+1.$$

The event $\{S_{\text{new}} < q_\alpha\}$ implies that the future score is ranked among the first $k_\alpha$ positions in the combined calibration-plus-future sample. Hence

$$\mathbb{P}\{S_{\text{new}} < q_\alpha \mid \widehat{\Delta}, \widehat{\sigma}, \text{nuisance fits}\} \le \frac{k_\alpha}{m+1} \le \alpha.$$

Taking expectation over the fitted functions gives

$$\mathbb{P}\{S_{\text{new}} \ge q_\alpha\} \ge 1 - \alpha.$$

Substituting the definition of $S_{\text{new}}$ yields

$$S_{\text{new}} \ge q_\alpha \iff \widehat{\Delta}(\boldsymbol{z}_{\text{new}}) - \Gamma_{\text{new}} \ge q_\alpha \iff \Gamma_{\text{new}} \le \widehat{\Delta}(\boldsymbol{z}_{\text{new}}) - q_\alpha = U_\alpha(\boldsymbol{z}_{\text{new}}).$$

Therefore

$$\mathbb{P}\{\Gamma_{\text{new}} \le U_\alpha(\boldsymbol{z}_{\text{new}})\} \ge 1 - \alpha.$$

For false-suppression control, define

$$F = \{h_\alpha^{\text{coco}}(\boldsymbol{z}_{\text{new}}) = 0,\ \Gamma_{\text{new}} \ge 0\}.$$

On $F$, the CoCo gate suppresses, so

$$h_\alpha^{\text{coco}}(\boldsymbol{z}_{\text{new}}) = 0 \implies U_\alpha(\boldsymbol{z}_{\text{new}}) < 0.$$

Together with $\Gamma_{\text{new}} \ge 0$, this gives

$$\Gamma_{\text{new}} > U_\alpha(\boldsymbol{z}_{\text{new}}).$$

Thus

$$F \subseteq \{\Gamma_{\text{new}} > U_\alpha(\boldsymbol{z}_{\text{new}})\}.$$

Using the coverage result,

$$\mathbb{P}(F) \le \mathbb{P}\{\Gamma_{\text{new}} > U_\alpha(\boldsymbol{z}_{\text{new}})\} \le \alpha.$$

This proves both claims.

### A.2 Proof of Corollary 1

For T-CoCo, the fitted function in Theorem 1 is

$$\widehat{\Delta} = \widehat{\Delta}_{\mathrm{T}}.$$

The calibration target is

$$\Gamma_i = \Gamma_i^{\mathrm{match}} = Y_i^{\mathrm{imp}} - Y_i^{\mathrm{sup}}.$$

Hence the calibration score is

$$S_i = \widehat{\Delta}_{\mathrm{T}}(\boldsymbol{z}_i) - \Gamma_i^{\mathrm{match}}.$$

The corollary assumes that these matched calibration scores and the future matched score are exchangeable after conditioning on the fitted T-learner. Therefore Assumptions 1–2 hold for the target $\Gamma^{\mathrm{match}}$. Applying Theorem 1 gives

$$\mathbb{P}\{h_\alpha^{\mathrm{coco}}(\boldsymbol{z}_{\mathrm{new}}) = 0, \Gamma_{\mathrm{new}}^{\mathrm{match}} \geq 0\} \leq \alpha.$$

If the matched construction is unbiased for the potential-outcome contrast in Eq. (1), then this target-level statement has the corresponding causal uplift interpretation.

### A.3 Proof of Proposition 1

Fix a value $\boldsymbol{z} = \boldsymbol{z}_0$ and suppress the argument $\boldsymbol{z}_0$ in the notation. Define

$$\mu_1 = \mathbb{E}[Y \mid H = 1, \boldsymbol{z}_0], \qquad \mu_0 = \mathbb{E}[Y \mid H = 0, \boldsymbol{z}_0], \qquad \eta = \mathbb{P}(H = 1 \mid \boldsymbol{z}_0).$$

By consistency and ignorability in Assumption 3,

$$\mu_1 = \mathbb{E}[Y_{h=1} \mid \boldsymbol{z}_0], \qquad \mu_0 = \mathbb{E}[Y_{h=0} \mid \boldsymbol{z}_0],$$

and therefore

$$\Delta(\boldsymbol{z}_0) = \mu_1 - \mu_0.$$

For arbitrary nuisance estimates $\widehat{\mu}_1$, $\widehat{\mu}_0$, and $\widehat{\eta}$, the DR pseudo-outcome is

$$\Phi = \widehat{\mu}_1 - \widehat{\mu}_0 + \frac{H}{\widehat{\eta}}(Y - \widehat{\mu}_1) - \frac{1-H}{1-\widehat{\eta}}(Y - \widehat{\mu}_0).$$

First suppose the propensity model is correct, so $\widehat{\eta} = \eta$. Then

$$\begin{aligned}
\mathbb{E}\left[\frac{H}{\eta}(Y - \widehat{\mu}_1) \mid \boldsymbol{z}_0\right] &= \frac{1}{\eta}\mathbb{E}[H(Y - \widehat{\mu}_1) \mid \boldsymbol{z}_0] \\
&= \frac{1}{\eta}\mathbb{P}(H = 1 \mid \boldsymbol{z}_0)\mathbb{E}[Y - \widehat{\mu}_1 \mid H = 1, \boldsymbol{z}_0] \\
&= \mu_1 - \widehat{\mu}_1.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\frac{1-H}{1-\eta}(Y - \widehat{\mu}_0) \mid \boldsymbol{z}_0\right] &= \frac{1}{1-\eta}\mathbb{P}(H = 0 \mid \boldsymbol{z}_0)\mathbb{E}[Y - \widehat{\mu}_0 \mid H = 0, \boldsymbol{z}_0] \\
&= \mu_0 - \widehat{\mu}_0.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}[\Phi \mid \boldsymbol{z}_0] &= \widehat{\mu}_1 - \widehat{\mu}_0 + (\mu_1 - \widehat{\mu}_1) - (\mu_0 - \widehat{\mu}_0) \\
&= \mu_1 - \mu_0 = \Delta(\boldsymbol{z}_0).
\end{aligned}$$

Second suppose the outcome nuisance models are correct, so $\widehat{\mu}_1 = \mu_1$ and $\widehat{\mu}_0 = \mu_0$, while $\widehat{\eta}$ may be misspecified. Then

$$\mathbb{E}[Y - \widehat{\mu}_1 \mid H = 1, \boldsymbol{z}_0] = 0, \qquad \mathbb{E}[Y - \widehat{\mu}_0 \mid H = 0, \boldsymbol{z}_0] = 0.$$

Hence both augmentation terms have conditional mean zero:

$$\mathbb{E}\left[\frac{H}{\widehat{\eta}}(Y - \widehat{\mu}_1) \mid \boldsymbol{z}_0\right] = 0, \qquad \mathbb{E}\left[\frac{1-H}{1-\widehat{\eta}}(Y - \widehat{\mu}_0) \mid \boldsymbol{z}_0\right] = 0.$$

Thus

$$\mathbb{E}[\Phi \mid \boldsymbol{z}_0] = \mu_1 - \mu_0 = \Delta(\boldsymbol{z}_0).$$

Since $\boldsymbol{z}_0$ was arbitrary, $\mathbb{E}[\Phi \mid \boldsymbol{z}] = \Delta(\boldsymbol{z})$.

### A.4 Proof of Corollary 2

Set

$$\widehat{\Delta} = \widehat{\Delta}_{\mathrm{DR}}, \qquad \Gamma_i = \Phi_i.$$

Then the calibration score is

$$S_i = \widehat{\Delta}_{\mathrm{DR}}(\boldsymbol{z}_i) - \Phi_i.$$

The corollary assumes that the calibration scores and the future score are exchangeable. Therefore Theorem 1 gives

$$\mathbb{P}\{h_\alpha^{\mathrm{coco}}(\boldsymbol{z}_{\mathrm{new}}) = 0, \Phi_{\mathrm{new}} \geq 0\} \leq \alpha.$$

This is a target-level statement for the DR pseudo-outcome. By Proposition 1, under Assumption 3 and either correct propensity or correct outcome nuisance models,

$$\mathbb{E}[\Phi \mid \boldsymbol{z}] = \Delta(\boldsymbol{z}).$$

Therefore the pseudo-outcome calibrated by DR-CoCo is conditionally mean-correct for the causal impression/suppression uplift.

### A.5 Proof of Proposition 2

For CoCo-M, fix a group value $g$. Let

$$\mathcal{I}_g = \{i \in \mathcal{D}_{\mathrm{cal}} : G(\boldsymbol{z}_i) = g\}$$

be the calibration indices in group $g$, and define

$$S_i = \widehat{\Delta}(\boldsymbol{z}_i) - \Gamma_i, \qquad i \in \mathcal{I}_g.$$

By the proposition assumption, conditional on $G = g$ and the fitted functions, these scores and the future score in the same group are exchangeable. Applying Theorem 1 within the group gives

$$\mathbb{P}\{h_\alpha^{\mathrm{coco}}(\boldsymbol{z}) = 0, \Gamma \geq 0 \mid G(\boldsymbol{z}) = g\} \leq \alpha.$$

This proves the conditional group validity. Marginal validity follows by averaging over $g$ whenever the group rule and any sparse-group fallback are fixed before evaluation.

For CoCo-F, condition on the fitted scale function $\widehat{\sigma}$ and assume $\widehat{\sigma}(\boldsymbol{z}) > 0$. Define normalized scores

$$S_i^{\mathrm{norm}} = \frac{\widehat{\Delta}(\boldsymbol{z}_i) - \Gamma_i}{\widehat{\sigma}(\boldsymbol{z}_i)}, \qquad S_{\mathrm{new}}^{\mathrm{norm}} = \frac{\widehat{\Delta}(\boldsymbol{z}_{\mathrm{new}}) - \Gamma_{\mathrm{new}}}{\widehat{\sigma}(\boldsymbol{z}_{\mathrm{new}})}.$$

By assumption, the normalized calibration scores and the future normalized score are exchangeable. Applying Theorem 1 to normalized scores gives

$$\mathbb{P}\{S_{\mathrm{new}}^{\mathrm{norm}} \geq q_{\alpha,\mathrm{norm}}\} \geq 1 - \alpha.$$

Since $\widehat{\sigma}(\boldsymbol{z}_{\mathrm{new}}) > 0$,

$$S_{\mathrm{new}}^{\mathrm{norm}} \geq q_{\alpha,\mathrm{norm}} \iff \Gamma_{\mathrm{new}} \leq \widehat{\Delta}(\boldsymbol{z}_{\mathrm{new}}) - q_{\alpha,\mathrm{norm}}\widehat{\sigma}(\boldsymbol{z}_{\mathrm{new}}).$$

Thus CoCo-F has the same upper-bound coverage and false-suppression control with

$$q_\alpha(\boldsymbol{z}) = q_{\alpha,\mathrm{norm}}\widehat{\sigma}(\boldsymbol{z}).$$

### A.6 Proof of Theorem 2

For each calibrated candidate $j$, let $F_j$ be the event that a nonnegative target is suppressed. Theorem 1 bounds $\mathbb{P}(F_j)$ by $\alpha_{\mathrm{item}}$. A union bound over at most $m_{\max}$ candidates gives $\mathbb{P}(\bigcup_j F_j) \leq m_{\max}\alpha_{\mathrm{item}}$, without independence between gates. The smaller number of displayed slots cannot replace the number of calibrated decisions.

### A.7 Proof of Proposition 3

Fix the gate $h_\alpha^{\mathrm{coco}}$ and policy $\pi_{\boldsymbol{\theta}}$ before evaluating the sample. For a candidate row, write

$$\kappa_\pi(\boldsymbol{z}) = \mathbf{1}\{L_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{x}, w) \neq \emptyset\}\lambda(L_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{x}, w)).$$

Let

$$m_1(\boldsymbol{z}) = \mathbb{E}[Y \mid A = 1, \boldsymbol{z}]$$

be the displayed-outcome regression target for ranking value. The item-level DR contribution in Eq. (4) is

$$Z_{\mathrm{DR}}(\pi) = \kappa_\pi(\boldsymbol{z})\widehat{m}_1(\boldsymbol{z}) + \frac{A}{\widehat{\omega}(\boldsymbol{z})}\kappa_\pi(\boldsymbol{z})\{Y - \widehat{m}_1(\boldsymbol{z})\}.$$

If $\widehat{\omega} = \omega$, then, conditional on $\boldsymbol{z}$,

$$\begin{aligned}\mathbb{E}\left[\frac{A}{\omega(\boldsymbol{z})}\kappa_\pi(\boldsymbol{z})\{Y - \widehat{m}_1(\boldsymbol{z})\} \mid \boldsymbol{z}\right] &= \kappa_\pi(\boldsymbol{z})\frac{1}{\omega(\boldsymbol{z})}\mathbb{E}[A\{Y - \widehat{m}_1(\boldsymbol{z})\} \mid \boldsymbol{z}] \\ &= \kappa_\pi(\boldsymbol{z})\mathbb{E}[Y - \widehat{m}_1(\boldsymbol{z}) \mid A = 1, \boldsymbol{z}] \\ &= \kappa_\pi(\boldsymbol{z})\{m_1(\boldsymbol{z}) - \widehat{m}_1(\boldsymbol{z})\}.\end{aligned}$$

Therefore

$$\mathbb{E}[Z_{\mathrm{DR}}(\pi) \mid \boldsymbol{z}] = \kappa_\pi(\boldsymbol{z})m_1(\boldsymbol{z}).$$

If instead $\widehat{m}_1 = m_1$, then

$$\mathbb{E}[Y - \widehat{m}_1(\boldsymbol{z}) \mid A = 1, \boldsymbol{z}] = 0,$$

so the augmentation term has conditional mean zero for any $\widehat{\omega}$ satisfying overlap. Hence again

$$\mathbb{E}[Z_{\mathrm{DR}}(\pi) \mid \boldsymbol{z}] = \kappa_\pi(\boldsymbol{z})m_1(\boldsymbol{z}).$$

Summing this equality over candidates in each request and then taking expectation over requests proves unbiasedness for $V(\pi_{\boldsymbol{\theta}})$.

### A.8 Proof of Theorem 3

For a fixed policy $\pi \in \Pi$, define the request-level IPS contribution

$$Z_i(\pi) = \sum_{w \in W_i} \frac{A_i(w)\kappa_\pi(\boldsymbol{x}_i, w)Y_i(w)}{\omega_i(w)}.$$

By the assumptions, at most $K$ terms can contribute, each outcome and location weight is in $[0, 1]$, and each nonzero inclusion probability is at least $\epsilon_\omega$. Therefore

$$0 \leq Z_i(\pi) \leq \frac{K}{\epsilon_\omega}.$$

The IPS estimator is

$$\widehat{V}_{\mathrm{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} Z_i(\pi).$$

For fixed $\pi$, Hoeffding's inequality gives, for any $u > 0$,

$$\mathbb{P}\{|\widehat{V}_{\mathrm{IPS}}(\pi) - V(\pi)| > u\} \leq 2\exp\left(-\frac{2nu^2\epsilon_\omega^2}{K^2}\right).$$

Taking a union bound over the finite class $\Pi$,

$$\mathbb{P}\left\{\sup_{\pi \in \Pi} |\widehat{V}_{\mathrm{IPS}}(\pi) - V(\pi)| > u\right\} \leq 2|\Pi|\exp\left(-\frac{2nu^2\epsilon_\omega^2}{K^2}\right).$$

Set the right-hand side equal to $\delta$ and solve for $u$:

$$u = \frac{K}{\epsilon_\omega}\sqrt{\frac{\log(2|\Pi|/\delta)}{2n}}.$$

This proves the uniform deviation statement.

Let

$$\pi^\star \in \arg\max_{\pi\in\Pi} V(\pi), \qquad \widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{V}_{\mathrm{IPS}}(\pi).$$

On the uniform event, for all $\pi \in \Pi$,

$$|\widehat{V}_{\mathrm{IPS}}(\pi) - V(\pi)| \le u.$$

Then

$$\begin{aligned} V(\pi^\star) - V(\widehat{\pi}) &\le V(\pi^\star) - \widehat{V}_{\mathrm{IPS}}(\pi^\star) + \widehat{V}_{\mathrm{IPS}}(\pi^\star) - \widehat{V}_{\mathrm{IPS}}(\widehat{\pi}) + \widehat{V}_{\mathrm{IPS}}(\widehat{\pi}) - V(\widehat{\pi}) \\ &\le u + 0 + u \\ &= 2u. \end{aligned}$$

Substituting the value of $u$ gives the result.

### A.9 Proof of Theorem 4

Condition on $\mathcal{F}_t$ as stated, retaining randomness in the calibration scores. Exchangeability of the $m_t + 1$ scores implies $\mathbb{P}\{S_t < S_{(k_t)} \mid \mathcal{F}_t\} \le k_t/(m_t + 1)$, with ties handled conservatively. The event $h_t^{\mathrm{coco}} = 0, \Gamma_t \ge 0$ implies $\Gamma_t > U_{\alpha,t}$ and hence $S_t < S_{(k_t)}$. This proves the one-step bound. Taking expectations and summing proves the average bound without independence between rounds. The floor and ceiling statements follow from their order-statistic indices. If exploration changes the gate only on an event of probability $\varepsilon_t$, a union bound adds at most $\varepsilon_t$ to the nominal-gate error probability. For CoCo-M, the same argument applies within each group. For CoCo-F, it applies to normalized scores.

Conditioning on the entire observed history would generally fix the calibration scores, destroying the rank-symmetry argument. The theorem therefore does not claim that arbitrary rolling buffers satisfy its hypothesis. Independent calibration after freezing a model is a sufficient split-conformal construction under exchangeable targets. Under distribution shift, additional procedures or explicit coverage-loss bounds are needed (Barber et al., 2023; Gibbs & Candès, 2021).

### A.10 Proof of Proposition 4

For each round, add and subtract the value of the best policy under the actual gated feasible set:

$$V_t(\pi_t^\star) - V_t(\pi_t) = \{V_t(\pi_t^\star) - V_t(\pi_t^{\mathrm{gate}\star})\} + \{V_t(\pi_t^{\mathrm{gate}\star}) - V_t(\pi_t)\}.$$

The second term sums to $\mathrm{Reg}_T^{\mathrm{base}}$. By the explicit value-loss assumption, the sum of the first terms is bounded by

$$B\sum_t (M_{\mathrm{FS},t} + M_{\mathrm{HI},t}) = BT(\hat{\mathcal{R}}_{\mathrm{FS},T} + \hat{\mathcal{R}}_{\mathrm{HI},T}),$$

which proves the claim.

A sufficient special case is a gate-sign oracle together with a per-candidate change in optimal value bounded by $b$ and at most $m_{\max}$ candidates: sequentially changing discrepant gate bits yields $B = bm_{\max}$. Arbitrary page interactions need not admit such an oracle or loss bound. These assumptions require justification beyond marginal uplift prediction. If values are averaged over random contexts and policies, applying the argument conditionally and then averaging gives the corresponding bound in expectation.

## B Algorithmic Details

CoCo training begins with separate model-fitting and calibration sets. After fitting the selected uplift estimator and specifying the calibration target $\Gamma_i$, the calibration scores determine the CoCo-1, CoCo-M, or CoCo-F correction. Offline CoCo-LTR then applies the gate before ranking and evaluates the resulting gated ranker with the benchmark, IPS, or DR value estimator. Online CoCo-LTR maintains a base online ranker, a rolling training window, and a rolling calibration buffer. At

round $t$, $U_{\alpha,t}$ is computed before serving, the base ranker is run only on eligible candidates, feedback and propensities are logged, and all models and calibration buffers are updated for round $t + 1$.

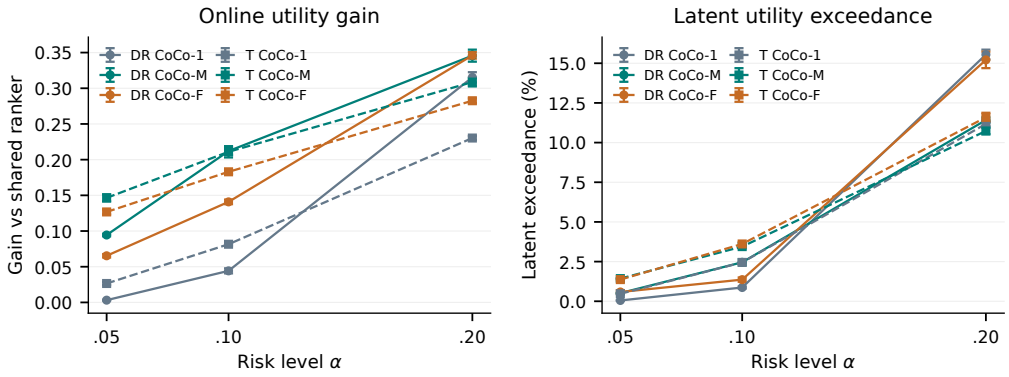


Figure 3: Online CoCo-LTR gain and latent exceedance as $\alpha$ varies. Exceedance denotes $\Delta > U_\alpha$, not noisy calibration-target coverage.

# C Experimental Details

This appendix describes data preparation, evaluation protocols, and sensitivity analyses.

## C.1 Common Protocol

Offline studies separate fitting, calibration, and evaluation data as detailed below. The online study evaluates each query before updating the models and rolling buffer. Its preprocessing is described in Appendix C.7. We report utility and suppression alongside the relevant exceedance metric. For offline tables, violation is measured against the experiment-specific calibration target $\Gamma_i$. The online study measures exceedance against latent utility $\Delta$. This metric is reported separately because it does not establish calibration-target coverage. The main alpha grid is $\alpha \in \{0.05, 0.10, 0.20\}$.

## C.2 Practical Construction of Logged Gate Data

CoCo requires data that distinguish intentional suppression from ordinary non-display. In production, this can be obtained with a small randomized eligibility bucket: for a controlled subset of traffic and candidates, the system randomizes $H$ before ranking, logs the gate propensity $\eta$, then logs whether the impressed candidate was actually displayed through $A$ and its ranking inclusion propensity $\omega$. A second common design is module-level or product-rule randomization, where an advertisement module, recommendation card, or answer module is randomly made eligible or suppressed for a small traffic slice. A third design uses historical rules only when the rule, its assignment probability, and the information available before assignment are logged well enough to define $\eta$.

A candidate that is not displayed because it ranked below the available slots has $A = 0$, but it does not necessarily have $H = 0$. Such rows are useful for ranking-exposure modeling through $\omega$, but they are not suppression-side examples for causal uplift unless the system intentionally suppressed them. The datasets provide different forms of feedback. Criteo supplies randomized treatment-control variation as a binary gate-arm analogue. MSLR provides full-information relevance-derived utility, and the online simulation includes controlled gate variation. Open Bandit provides randomized action-level propensities for off-policy evaluation.

### C.3 Dataset Roles

Table 3: Empirical roles of the datasets.

| Dataset | Model family | Decision comparison |
|---|---|---|
| Criteo Uplift v2.1 | LightGBM T/DR learners | Same uplift model with point gate versus CoCo-1/M/F |
| MSLR-WEB10K offline | LambdaRank-family rankers plus T/DR gates | Same trained top-$K$ ranker with and without CoCo-1/M/F |
| Synthetic online LTR | Timestamp-updated ranker and T/DR gates | Same online ranker with point, CoCo-1, CoCo-M, and CoCo-F decision layers |
| Open Bandit | LightGBM action/click score model | Supporting logged-bandit IPS for cost-sensitive action-level gates |

### C.4 Criteo Details

The Criteo study uses Criteo Uplift v2.1 randomized treatment/control data with conversion as the outcome. The raw file is sampled randomly across the full gzip before splitting. The experiment uses a reservoir of 2M rows, seeds 2026–2030, and 400k rows per seed. Each seed uses 50% training, 20% calibration, and 30% test. We compare LightGBM T-learners and DR-learners with point and CoCo gates. Because Criteo has no finite-slot top-$K$ ranking stage, the comparison excludes a keep-all/display-all ranking baseline.

Table 4: Criteo point-gate comparison at $\alpha = 0.10$. A keep-all/display-all ranking baseline is omitted because Criteo has no finite-slot top-$K$ ranking stage. Values are reported in units of $10^{-4}$, and rate columns are percentages.

| Learner | Decision | Value / $10^{-4}$ | Gain vs point (%) | Violation (%) | Suppression (%) |
|---|---|---|---|---|---|
| DR | Point | 8.75 | 0.00 | – | 12.6 |
| DR | CoCo-1 | 8.83 | 0.98 | 10.1 | 7.7 |
| DR | CoCo-M | 8.79 | 0.44 | 10.0 | 15.3 |
| **DR** | **CoCo-F** | **8.93** | **2.04** | **10.0** | **7.1** |
| T | Point | 9.97 | 0.00 | – | 1.0 |
| T | CoCo-1 | 9.90 | -0.79 | 9.7 | 9.4 |
| T | CoCo-M | 9.86 | -1.18 | 9.6 | 9.4 |
| T | CoCo-F | 9.96 | -0.14 | 9.5 | 9.3 |

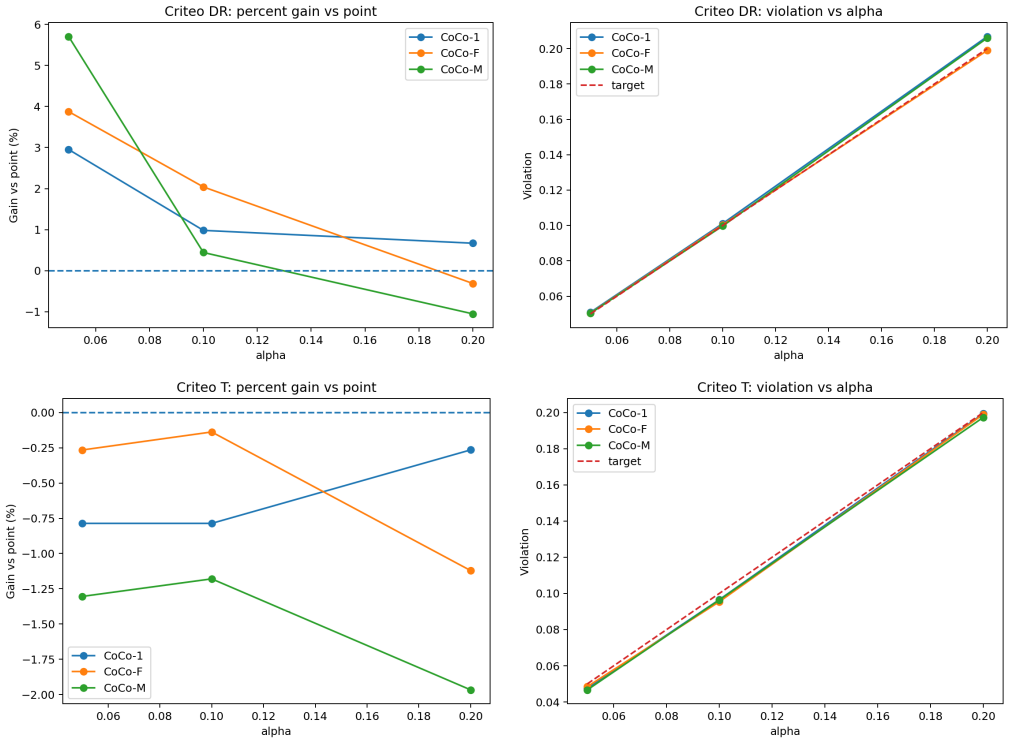


Figure 4: Criteo point-gate comparison across risk levels. The top and bottom rows show DR and T learners, respectively. The left column shows value gain, and the right column shows empirical target violation.

### C.5 MSLR-WEB10K Details and Sensitivity

The MSLR experiment uses the official Fold1 split: train on Fold1 `train.txt`, calibrate on `vali.txt`, and evaluate on `test.txt`. It uses 723,412 training documents, 235,259 calibration documents, and 241,521 test documents. The cost-grid analysis reports both T-style and DR-style gate learners with CoCo-1, CoCo-M, and CoCo-F at $\alpha \in \{0.05, 0.10, 0.20\}$ and display costs $0.05$, $0.10$, $0.15$, $0.20$, and $0.25$.

Table 5 summarizes the constructed net-utility target across display costs. Since normalized gain is $(2^{\text{rel}} - 1)/15$, the set of nonnegative relevance grades changes only when the display cost crosses a relevance-gain threshold. The target is substantially less suppression-dominated at cost $0.10$ than at cost $0.25$, which motivates using cost $0.10$ in the main MSLR table.

Table 5: MSLR Fold1 test target-level diagnostic across display costs. The main MSLR table uses cost $0.10$, where the ideal top-10 target is not dominated by negative-net-utility documents. Cost $0.25$ is included as a high-cost sensitivity case in which many even ideally ranked top-10 candidates are negative-net-utility.

| Cost | All docs neg. (%) | Ideal top-10 neg. (%) | Queries with $< 10$ nonneg. docs (%) | Ideal top-10 utility |
|---|---|---|---|---|
| 0.02 | 51.7 | 8.7 | 13.0 | 1.382 |
| 0.05 | 51.7 | 8.7 | 13.0 | 1.246 |
| 0.10 | 83.9 | 23.7 | 36.4 | 1.021 |
| 0.15 | 83.9 | 23.7 | 36.4 | 0.795 |
| 0.20 | 83.9 | 23.7 | 36.4 | 0.569 |
| 0.25 | 97.4 | 72.7 | 92.4 | 0.344 |

Table 6: Model-level MSLR cost grid for the net ranker with the DR gate learner at $\alpha = 0.10$. Suppression and violation are percentages. At cost $0.10$, CoCo-M improves utility by $+0.0295$ with $39.4\%$ suppression.

| Cost | Ranker $U$ | CoCo-F $\Delta U$ / Supp. | CoCo-M $\Delta U$ / Supp. | CoCo-M viol. (%) |
|---|---|---|---|---|
| 0.05 | 0.4821 | +0.0002 / 2.2 | +0.0001 / 0.9 | 9.8 |
| **0.10** | **0.2553** | **+0.0155 / 24.4** | **+0.0295 / 39.4** | **9.8** |
| **0.15** | **0.0312** | **+0.0599 / 43.8** | **+0.0514 / 39.4** | **9.8** |
| **0.20** | **−0.2203** | **+0.1581 / 64.4** | **+0.1485 / 65.5** | **9.8** |
| **0.25** | **−0.4445** | **+0.3152 / 86.8** | **+0.4563 / 96.7** | **9.8** |

Table 7: Full MSLR conformal ablation at display cost $0.10$ and $\alpha = 0.10$. Values are means over three seeds with standard errors. Rate columns are percentages.

| Ranker | Gate learner | Decision | Utility@10 | Gain vs ranker | Violation (%) | Suppression (%) |
|---|---|---|---|---|---|---|
| net | DR | CoCo-1 | $0.2553 \pm 0.0001$ | $0.0000$ | 9.8 | 0.0 |
| **net** | **DR** | **CoCo-M** | $\mathbf{0.2848 \pm 0.0010}$ | $\mathbf{+0.0295 \pm 0.0009}$ | **9.8** | **39.4** |
| net | DR | CoCo-F | $0.2708 \pm 0.0005$ | $+0.0155 \pm 0.0003$ | 9.8 | 24.4 |
| net | DR | point diag. | $0.3104 \pm 0.0013$ | $+0.0551 \pm 0.0012$ | – | 89.7 |
| net | DR | ranker | $0.2553 \pm 0.0001$ | $0.0000$ | – | 0.0 |
| net | T | CoCo-1 | $0.2553 \pm 0.0001$ | $0.0000$ | 9.8 | 0.0 |
| **net** | **T** | **CoCo-M** | $\mathbf{0.2846 \pm 0.0009}$ | $\mathbf{+0.0293 \pm 0.0008}$ | **9.9** | **39.4** |
| net | T | CoCo-F | $0.2707 \pm 0.0002$ | $+0.0154 \pm 0.0002$ | 9.7 | 24.5 |
| net | T | point diag. | $0.3128 \pm 0.0007$ | $+0.0576 \pm 0.0009$ | – | 89.2 |
| net | T | ranker | $0.2553 \pm 0.0001$ | $0.0000$ | – | 0.0 |
| rel | DR | CoCo-1 | $0.2521 \pm 0.0004$ | $0.0000$ | 9.8 | 0.0 |
| **rel** | **DR** | **CoCo-M** | $\mathbf{0.2802 \pm 0.0004}$ | $\mathbf{+0.0281 \pm 0.0008}$ | **9.8** | **39.4** |
| rel | DR | CoCo-F | $0.2673 \pm 0.0006$ | $+0.0152 \pm 0.0004$ | 9.8 | 24.4 |
| rel | DR | point diag. | $0.3041 \pm 0.0014$ | $+0.0520 \pm 0.0019$ | – | 89.7 |
| rel | DR | ranker | $0.2521 \pm 0.0004$ | $0.0000$ | – | 0.0 |
| rel | T | CoCo-1 | $0.2521 \pm 0.0004$ | $0.0000$ | 9.8 | 0.0 |
| **rel** | **T** | **CoCo-M** | $\mathbf{0.2804 \pm 0.0002}$ | $\mathbf{+0.0283 \pm 0.0003}$ | **9.9** | **39.4** |
| rel | T | CoCo-F | $0.2674 \pm 0.0004$ | $+0.0154 \pm 0.0001$ | 9.7 | 24.5 |
| rel | T | point diag. | $0.3052 \pm 0.0005$ | $+0.0531 \pm 0.0003$ | – | 89.2 |
| rel | T | ranker | $0.2521 \pm 0.0004$ | $0.0000$ | – | 0.0 |

Table 8: MSLR net-ranker DR sensitivity at display cost $0.10$. Larger $\alpha$ permits more aggressive suppression and higher utility, while empirical violation tracks the target level.

| Decision | $\alpha$ | Utility@10 | Gain vs ranker | Violation (%) | Suppression (%) |
|---|---|---|---|---|---|
| CoCo-1 | 0.05 | $0.2553 \pm 0.0001$ | $0.0000$ | 5.0 | 0.0 |
| CoCo-M | 0.05 | $0.2662 \pm 0.0003$ | $+0.0110 \pm 0.0002$ | 4.9 | 19.6 |
| CoCo-F | 0.05 | $0.2600 \pm 0.0003$ | $+0.0047 \pm 0.0001$ | 5.0 | 12.4 |
| CoCo-1 | 0.10 | $0.2553 \pm 0.0001$ | $0.0000$ | 9.8 | 0.0 |
| CoCo-M | 0.10 | $0.2848 \pm 0.0010$ | $+0.0295 \pm 0.0009$ | 9.8 | 39.4 |
| CoCo-F | 0.10 | $0.2708 \pm 0.0005$ | $+0.0155 \pm 0.0003$ | 9.8 | 24.4 |
| CoCo-1 | 0.20 | $0.3045 \pm 0.0004$ | $+0.0492 \pm 0.0005$ | 19.7 | 64.7 |
| **CoCo-M** | **0.20** | $\mathbf{0.3075 \pm 0.0044}$ | $\mathbf{+0.0522 \pm 0.0044}$ | **19.6** | **72.2** |
| CoCo-F | 0.20 | $0.2948 \pm 0.0006$ | $+0.0395 \pm 0.0005$ | 19.6 | 49.9 |

### C.6 Open Bandit Supporting OPE Details

Open Bandit uses random-policy logs from the all, women, and men campaigns. Each logged row contains a context, a randomly selected item action, its logging probability, and click feedback. The experiment trains an action-level score model, applies a cost-sensitive point or CoCo gate to the deterministic target action, and evaluates the resulting target policy by IPS. This experiment evaluates the contextual-bandit special case of the gate. It has no top-$K$ assignment stage or separate eligibility and slot-assignment decisions. The IPS estimate assigns nonzero weight only when the

deterministic target action matches the logged random action, limiting the comparison to action-level off-policy evaluation.

Table 9: Open Bandit point-versus-CoCo cost sensitivity at $\alpha = 0.10$ for the displayed costs. Point uses the same action-score model with threshold zero. Rate columns are percentages.

| Cost | Decision | IPS net | Gain vs vanilla (%) | Violation (%) | Suppression (%) |
|---|---|---|---|---|---|
| 0.005 | Point | 0.0487 | 0.0 | – | 6.7 |
| **0.005** | **CoCo-1** | **0.0492** | **1.0** | **7.5** | **31.5** |
| 0.005 | CoCo-M | 0.0481 | -1.3 | 5.5 | 34.8 |
| 0.005 | CoCo-F | 0.0487 | 0.1 | 9.2 | 43.2 |
| 0.010 | Point | 0.0457 | 5.1 | – | 32.5 |
| 0.010 | CoCo-1 | 0.0461 | 6.0 | 7.5 | 46.8 |
| 0.010 | CoCo-M | 0.0450 | 3.5 | 5.5 | 48.9 |
| **0.010** | **CoCo-F** | **0.0463** | **6.5** | **7.5** | **63.4** |
| 0.050 | Point | 0.0333 | 1594.6 | – | 83.3 |
| 0.050 | CoCo-1 | 0.0338 | 1618.1 | 20.2 | 87.5 |
| 0.050 | CoCo-M | 0.0325 | 1553.6 | 5.5 | 78.7 |
| **0.050** | **CoCo-F** | **0.0343** | **1644.9** | **20.2** | **89.0** |

### C.7 Sequential Online Simulation Details

The sequential simulation uses seeds 2026–2028, 900 queries with 24 candidates each, and $K = 10$. There are 120 warm-up queries and 120 additional queries to populate the buffer. Reported metrics average query indices 240–899, giving 660 evaluated queries per seed. A regime change at index 540 alters latent utility and noise. Contexts and item features are Gaussian. Nonlinear terms, three context groups, and heteroscedastic noise introduce variation in candidate utility and prediction difficulty.

The common ranker is an online linear squared-loss model on 19 enriched features. All gates use the same pre-update ranking scores at each query. The ranker is then updated using full-information relevance feedback for all candidates. The T-learner uses two outcome regressions and calibrates to $y_1 - y_0$. The DR learner regresses a pseudo-outcome constructed from randomized gate feedback with known propensities clipped by design to $[0.20, 0.80]$. Both maintain an online residual-scale regressor. The calibration buffer stores 2,880 candidate scores from past pre-update predictions and targets. Old scores are discarded as new scores arrive, and retained scores are not recomputed under the current model. CoCo-M uses the three context groups with a minimum of 30 calibration examples and a global fallback. These operations occur after evaluating each query.

Feature standardization uses the full generated stream, group thresholds use stream-level feature quantiles, and the regime indicator is an observed feature. Thus no future outcomes enter per-query parameter updates, but preprocessing is transductive in the covariates and the shift is partly observed. The implementation uses $k = \lceil (m+1)\alpha \rceil$ rather than the floor convention in Section 4.1. Even under exchangeability this has the finite-grid bound $k/(m+1)$, up to one order-statistic step above $\alpha$. Cached scores from evolving models and grouped candidates need not satisfy the theorem's exchangeability assumption.

The online exceedance metric computes $\mathbf{1}\{\Delta > U_\alpha\}$ using latent synthetic utility. It does not compute $\mathbf{1}\{\Gamma > U_\alpha\}$ for the T/DR calibration targets. We refer to this metric as *latent exceedance*. It upper-bounds false suppression relative to nonnegative latent utility on the evaluated sample, since $\{U_\alpha < 0, \Delta \geq 0\} \subseteq \{\Delta > U_\alpha\}$, but is not an empirical check of noisy-target conformal coverage. Target coverage would require evaluation against the corresponding T/DR calibration targets, which is not available from the recorded metrics. Online suppression is measured among the common ungated top-10. Eligible candidates below those positions may refill the vacated slots.

Means and SEs are computed from the three seed-level averages. Figure 5 cumulatively sums each seed's per-query utility difference from the shared ungated ranker and then plots the mean with one-SE bands. These bands describe variation across three simulations and are not simultaneous confidence bands. Table 11 compares CoCo-M with the point gate at a larger risk budget.

Table 10: Sequential online simulation details at $\alpha = 0.10$. Rate columns are percentages.

| Learner | Variant | Utility/query | Gain/query | Exceed. (%) | Suppression (%) |
|---|---|---|---|---|---|
| DR | CoCo-1 | 0.226 | +0.044 | 0.87 | 2.0 |
| **DR** | **CoCo-M** | **0.394** | **+0.212** | **2.46** | **9.7** |
| DR | CoCo-F | 0.323 | +0.141 | 1.36 | 6.3 |
| T | CoCo-1 | 0.263 | +0.082 | 2.45 | 3.6 |
| **T** | **CoCo-M** | **0.393** | **+0.211** | **3.44** | **9.8** |
| T | CoCo-F | 0.365 | +0.183 | 3.61 | 8.7 |

Table 11: Online comparison with the point gate at a larger risk budget. The point gate thresholds the same fitted online score at zero. CoCo-M uses the same stream and ranker with the indicated risk budget. Rate columns are percentages.

| Learner | Decision | $\alpha$ | Utility/query | Difference vs point | Exceed. (%) | Suppression (%) |
|---|---|---|---|---|---|---|
| DR | Point | – | 0.6315 | 0.0000 | – | 47.1 |
| **DR** | **CoCo-M** | **0.40** | **0.6651** | **+0.0335** | **38.2** | **38.0** |
| T | Point | – | 0.6393 | 0.0000 | – | 46.0 |
| **T** | **CoCo-M** | **0.40** | **0.6667** | **+0.0275** | **35.8** | **36.5** |

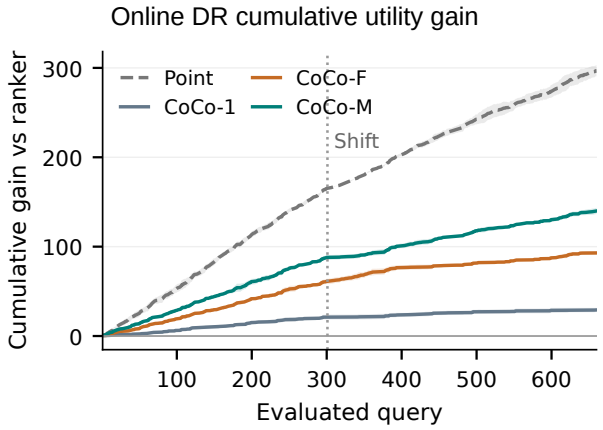


Figure 5: Online DR cumulative utility gains relative to the shared ungated ranker. Shaded bands show $\pm 1$ SE across three seeds, with $\alpha = 0.10$ for conformal gates. The dotted line marks the distribution shift. Point achieves higher cumulative utility than each CoCo variant in this experiment.