SKIP, ADAPT, OR ROUTE? CERTIFIED DEPLOYMENT OF TEST-TIME ADAPTIVE RERANKERS UNDER LIMITED LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation of a re-ranker's query representation from pseudo-relevance feedback improves some queries and damages others, so its average effect does not decide whether to deploy it. We treat this decision as a choice among frozen actions, made from a small labeled sample of the target traffic: keep the incumbent ranking, adapt every query, or route adaptation with a label-free gate. We prove three results. First, a selector that sees only label-free signals has regret bounded away from zero on conditions it cannot distinguish, so target labels are necessary. Second, an exact value ladder bounds what any gate on a signal z can add over the better simple action by $\min(\mathbb{E}[\mu(z)_+], \mathbb{E}[\mu(z)_-])$, where $\mu(z)$ is the conditional effect of adaptation; four anchor regimes determine its sign. Third, certified policy selection (CPS) replaces the incumbent only when a confidence bound on the paired improvement clears a margin. It has a finite-sample safety guarantee, an explicit regret rate that includes the cost of fitting a gate, a closed-form label budget, and a sequential variant. We replay stored per-query outputs of 23 benchmark conditions, a production search log, and a new self-adaptation testbed. With 50 labeled queries, uncertified selection deploys a harmful action in 23–70% of calibrations and an uncorrected t-test in 6–28%; CPS with a normal bound does so in under 1% on average, and with a finite-sample bound never, at the price of deploying rarely. The label budget is predictable from the effect size. Score entropy sees at most about a fifth of the per-query routing headroom, because adaptation mainly pulls the scorer toward the retriever, and recommendations from a fixed configuration reverse once step sizes and the retriever's own ranking enter the action set.

1 INTRODUCTION

Test-time adaptation (TTA) lets a neural re-ranker adjust itself to each query at inference time. A common form updates the query representation for a few gradient steps so that the top retrieved documents, used as pseudo-relevance feedback (PRF) anchors, score higher than the rest of the candidate pool (Rocchio, 1971; Yu et al., 2021; Sun et al., 2020). Such updates repair some queries and damage others. The average gain over a benchmark therefore does not answer the practical question: should a system adapt every query, adapt none, or *route* adaptation to the queries a cheap gate selects? And which of these should replace the ranking that is currently deployed?

The answer is less stable than it looks. Consider Quora duplicate-question retrieval with a BERT-large scorer re-ranking the top 50 candidates of a BGE-small retriever. Under a fixed step size, adapting every query loses 0.006 MRR@10 against the frozen scorer, so a fixed-configuration evaluation recommends SKIP. With a four-times smaller step and two anchors instead of three, adaptation *gains* 0.098, and the recommendation flips to ALL. Yet the retriever's own order scores 0.880, about 0.19 above every adapted re-ranker in the set, so the right deployment is not to re-rank at all (Table 1). Each recommendation is correct for its action set, and each set is a choice the evaluator made. A deployment method must therefore make the action set and the incumbent explicit. It must also say how many labeled target queries are needed before replacing the incumbent is safe.

We treat deployment as selection among a finite set of frozen actions using a small labeled calibration sample from the target traffic. Our contributions are:

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.

- **Labels are necessary (Proposition 4).** Any selector that sees only label-free signals, such as score entropy, anchor losses or a classifier trained on other conditions, has regret bounded away from zero on some pair of conditions that it cannot tell apart. Empirically, an entropy rule that works well against the frozen scorer replaces the retriever's order with a worse ranking in 14 of 19 conditions, and decisions transferred between conditions are wrong in 30 of 56 cases.
- **What routing can buy (Propositions 2–3).** An exact value ladder bounds the improvement of any gate on a signal z over the better simple action by $H_z = \min(\mathbb{E}[\mu(z)_+], \mathbb{E}[\mu(z)_-])$, where $\mu(z) = \mathbb{E}[\Delta \mid z]$ is the conditional effect of adaptation. Four anchor regimes determine the sign of μ, and a condition's majority regime predicts whether ALL or SKIP is better in all 17 conditions where it is regime I or II. Measured on 19 conditions, score entropy sees at most about a fifth of the per-query headroom, and a gate beats the better simple action by more than 0.005 MRR@10 in only 3. The mechanism is simple: PRF adaptation pulls the scorer toward the retriever, so it pays where the retriever is right and the scorer wrong, which the scorer's score shape barely reveals.
- **Certified policy selection (Section 4).** CPS replaces the incumbent only when a confidence bound on the paired improvement exceeds a margin. It has a finite-sample safety guarantee, an explicit regret rate, a bound on the cost of fitting a gate (Theorem 1, Lemma 1), a closed-form label budget (Corollary 1), and a time-uniform variant for sequential labeling.
- **A replay study.** We replay stored per-query outputs of 23 benchmark conditions, a production search log, and a new testbed in which the retriever's own encoder adapts its queries (10 datasets, 24 configurations each). With 50 labeled queries, the uncertified empirical best deploys a harmful action in 23–70% of calibrations, and an uncorrected paired t-test in 6–28%. CPS with a normal bound does so in at most 0.7% on average (worst condition 5%); the finite-sample variants never do, but rarely deploy at this n. Given the effect size, the label budget is predicted within a median 6% error.

We do not propose a new adaptation method; all re-rankers we study are weaker than their first stage. The contribution is a deployment procedure with guarantees and an account of when selective adaptation can pay.

2 PROBLEM SETUP

Pipeline and actions. A *condition* c fixes a query distribution P, a first-stage retriever that returns a candidate pool $\mathcal{D}_q$ for each query q, a scorer, and a relevance-labeling protocol. A finite *action set* $\mathcal{A}$ contains every ranking procedure under consideration. Our action sets include UPSTREAM (return the retriever's order of $\mathcal{D}_q$), SKIP (re-rank $\mathcal{D}_q$ with the frozen scorer), ALL_ϕ (adapt the query representation on every query with pseudo-relevance-feedback (PRF) anchors and hyperparameters ϕ; see "Adaptation" below), and $\text{ROUTE}_g(s, a)$, which applies action a on queries where a gate g fires and action s otherwise. One action $b \in \mathcal{A}$ is the *incumbent*: the system currently deployed. We make the incumbent explicit because the natural reference changes the conclusion: in the audit of Section 5.2, the retriever's own ranking beats both the frozen and the adapted scorer in most conditions.

Utilities and values. Fix a cutoff $k = 10$. For a per-query effectiveness measure (reciprocal rank, average precision or recall at k), $U_a(q) \in [0, 1]$ is the utility of action a on query q, computed from relevance labels. The value of a is $V(a) = \mathbb{E}_{q\sim P}[U_a(q)]$, and the regret of a deployed action $\hat{a}$ is $\mathcal{R}(\hat{a}) = \max_{a\in\mathcal{A}} V(a) - V(\hat{a})$. The maximum is over the stated action set only; it is neither a per-query oracle nor an optimum over all possible adaptation procedures. For two actions s, a we write the per-query difference $\Delta(q) = U_a(q) - U_s(q) \in [-1, 1]$, and for a candidate c its gap to the incumbent $\theta_c = V(c) - V(b)$.

Label-free signals and gates. A *label-free signal* $z(q) \in \mathcal{Z}$ is any statistic computable at serving time without relevance labels, e.g. the entropy or top-gap of the pre-update score distribution, or the pseudo-relevance loss before adaptation. A gate is a map $g : \mathcal{Z} \to \{0, 1\}$; then $U_{\text{ROUTE}_g(s,a)}(q) = U_s(q) + g(z(q))\,\Delta(q)$. We write $\mu(z) = \mathbb{E}[\Delta \mid z]$ for the conditional mean difference.

What calibration observes. A calibration sample is n queries drawn i.i.d. from P together with relevance labels for the documents that the actions in $\mathcal{A}$ place in the top k. Every action is a deterministic function of the query and pool, so the outcomes of *all* actions on a calibration query are computed offline; no counterfactual is missing. The scarce resource is relevance labeling, needed

only for documents some action ranks in its top k (at most the pool: 50 documents on BEIR, 20.7 on average on Qilin). The question is: given n labeled queries, which action should replace the incumbent, if any, and with what guarantee?

Two decision levels. The *scorer-level* decision fixes $b = \text{SKIP}$, as fixed-configuration benchmarks do; the *system-level* decision uses the deployed ranking, here UPSTREAM. Both are the same selection problem with different $(\mathcal{A}, b)$.

Adaptation. The adapted actions follow a standard PRF recipe. The anchors $\mathcal{P}_q$ are the top m candidates of the retriever. A query offset ω, initialized at zero with all model weights frozen, takes T gradient steps of size η on

$$\mathcal{J}_q(\omega) = -\log \frac{\sum_{i\in\mathcal{P}_q} \exp\big(f_i(\omega)/\tau\big)}{\sum_{j\in\mathcal{M}_q\cup\mathcal{P}_q} \exp\big(f_j(\omega)/\tau\big)} + \beta \sum_{j\in\mathcal{M}_q} \text{softplus}\big(\xi + f_j(\omega)\big) + \kappa\|\omega\|^2, \qquad (1)$$

where $f_j(\omega)$ scores candidate j against the offset query, $\mathcal{M}_q$ are the 20 highest-scoring non-anchors, and the middle term pushes them beyond distance ξ. The pool is then re-ranked with the offset query. In all our runs, scores are negative Poincaré-ball distances, $\tau = 0.05$, $\beta = 0.1$, $\xi = 0.5$ and $\kappa = 0.01$. We write ALL_ϕ with $\phi = (m, T, \eta)$. Appendix B gives the remaining terms and implementation details. The theory and CPS use only per-query utilities, so they apply to any adaptation procedure; the empirical findings are specific to this recipe.

3 WHEN CAN ROUTING PAY?

Fix a static action s and an adapted action a with difference $\Delta = U_a - U_s$. Write $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$, so that $G = \mathbb{E}[\Delta_+]$ is the total gain available from adaptation and $L = \mathbb{E}[\Delta_-]$ is the total harm. For a gate g, $G_g = \mathbb{E}[g(z)\Delta_+]$ and $L_g = \mathbb{E}[g(z)\Delta_-]$ are the gain and harm it admits. All proofs are in Appendix A; the propositions require only that utilities are bounded.

Proposition 1 (Exact decomposition). *$V(\text{ROUTE}_g) - V(s) = G_g - L_g$ and $V(\text{ROUTE}_g) - V(a) = (L - L_g) - (G - G_g)$. Hence* ROUTE *strictly improves on both simple actions if and only if the admitted gain exceeds the admitted harm and the avoided harm exceeds the forgone gain.*

Proposition 1 is an identity: it says what calibration must establish but not whether any label-free gate can establish it. The next result bounds what a gate can achieve.

Proposition 2 (Routing value ladder). *Let $\Gamma(g) = V(\text{ROUTE}_g) - \max\{V(s), V(a)\}$ be the improvement of a gate over the better simple action.*

(i) Over all gates that may depend on the query (including its labels), $\sup_g \Gamma(g) = H := \min(G, L)$.

(ii) Over gates that depend on q only through z, $\sup_g \Gamma(g) = H_z := \min\big(\mathbb{E}[\mu(z)_+], \mathbb{E}[\mu(z)_-]\big)$, attained by $g^\star(z) = \mathbb{1}[\mu(z) > 0]$.

(iii) $0 \le H_z \le H$; $H_z = H$ whenever, for almost every z, Δ does not take both signs given z; and $H_z = 0$ if and only if $\mu(z)$ has almost surely constant sign.

In treatment-choice terms (Manski, 2004; Kitagawa & Tetenov, 2018), μ is the conditional effect of "treating" a query with adaptation and H_z the advantage of the first-best rule on z over uniform rules; unlike in causal inference, both outcomes are observed, so H is identifiable. The ladder $H \ge H_z \ge \Gamma(\hat{g})$ separates the headroom the signal cannot see ($H - H_z$) from the price of estimating the gate ($H_z - \Gamma(\hat{g})$, Lemma 1). A large H with $H_z \approx 0$ means that z cannot tell helped from hurt queries, whatever the gate's form.

What must z reveal for $H_z > 0$? A two-state model of PRF anchors makes this explicit. Let $Y(q) \in \{0, 1\}$ indicate that the anchor set is informative (here: at least half of the anchors are relevant), and define the anchor precision $\rho(z) = \Pr(Y = 1 \mid z)$, the effect when anchors are informative $\gamma(z) = \mathbb{E}[\Delta \mid z, Y = 1]$, and the loss when they are not $\lambda(z) = -\mathbb{E}[\Delta \mid z, Y = 0]$.

Proposition 3 (Anchor regimes). *Assume $0 < \rho(z) < 1$. Then $\mu(z) = \rho(z)\gamma(z) - (1 - \rho(z))\lambda(z)$. At each z with $(\gamma, \lambda) \ne (0, 0)$ exactly one of four regimes holds (if $\gamma = \lambda = 0$, then $\mu = 0$):*

I. $\gamma > 0 \ge \lambda$ (adaptation helps whether or not the anchors are informative): $\mu > 0$;

Algorithm 1 Certified Policy Selection (CPS)

Input: action set $\mathcal{A}$ with incumbent b; route pairs $\{(s_j, a_j)\}_{j \le J}$; signal z and label-free threshold grid $\mathcal{T}$; labeled calibration queries $q_1, \dots, q_n$; risk level δ; minimum improvement $\epsilon \ge 0$; lower-bound routine LB

Randomly split the indices into halves I_{fit} and I_{cert}
for $j = 1, \dots, J$ **do**
    $\hat{g}_j \leftarrow \arg\max_g \sum_{i \in I_{\text{fit}}} g(z(q_i))\, \Delta_j(q_i)$ over gates $g(z) = \mathbb{1}[\pm(z - t) > 0], t \in \mathcal{T}$
$\mathcal{C} \leftarrow (\mathcal{A} \setminus \{b\}) \cup \{\textsc{Route}_{\hat{g}_j}(s_j, a_j)\}_j, \quad K \leftarrow |\mathcal{C}|$
for each candidate $c \in \mathcal{C}$ **do**
    $I_c \leftarrow I_{\text{cert}}$ if c is a Route action, else $\{1, \dots, n\}$
    $\ell_c \leftarrow \text{LB}\big(\{U_c(q_i) - U_b(q_i)\}_{i \in I_c},\ \delta/K\big)$
$\hat{c} \leftarrow \arg\max_c \ell_c$
return $\hat{c}$ if $\ell_{\hat{c}} > \epsilon$, otherwise b ▷ defer: keep the incumbent

II. $\gamma \le 0 < \lambda$ *(it hurts either way):* $\mu < 0$*;*

III. $\gamma > 0,\ \lambda > 0$*:* $\mu > 0$ *if and only if* $\rho > \rho^\star := \lambda/(\gamma + \lambda)$*;*

IV. $\gamma \le 0,\ \lambda \le 0$*:* $\mu > 0$ *if and only if* $\rho < \rho^\star$*.*

Consequently, $H_z > 0$ *if and only if* $\mu > 0$ *and* $\mu < 0$ *each hold on sets of* z *of positive probability, where the sign of* μ *is fixed by the regime and, in regimes III and IV, by the side of* $\rho^\star$*. In particular, if* (ρ, γ, λ) *does not vary with* z*, no gate on* z *can pay, however accurate* z *is as a difficulty or confidence signal.*

Anchor quality matters through ρ versus $\rho^\star$ (regime III) and is irrelevant in regimes I and II. Score shape matters only insofar as it *stratifies* queries across regimes or across $\rho^\star$: a sharp score distribution can indicate a correct top candidate or a confidently wrong one. The regime predicts the optimal action well (Section 5.5), but estimating it requires labels.

This last point holds in general: deployment cannot be decided without labels from the target condition. Call a *label-free observation* any random element O computed from queries, candidate pools and the outputs of all actions, but not from relevance labels. This includes every score-shape summary, every anchor-loss statistic and every input of a selector trained on other conditions.

Proposition 4 (Label-free selection is not identifiable)**.** *Let* $a_1, a_2 \in \mathcal{A}$ *produce different top-*k *lists on a set of queries of positive probability, and let the utility be reciprocal rank or average precision at* k*. There exist two conditions with the same distribution of* O *such that* $V(a_1) - V(a_2) = d_1 > 0$ *in the first and* $V(a_2) - V(a_1) = d_2 > 0$ *in the second. Every selector* $\psi(O) \in \{a_1, a_2\}$*, randomized or not, has regret at least* $d_1 d_2/(d_1 + d_2) \ge \min(d_1, d_2)/2$ *in one of the two conditions.*

The construction keeps queries, pools and scores fixed and moves relevance between the first two documents on which the lists disagree. The statement is worst-case: label-free rules can work (an entropy threshold separates ALL from SKIP well on Testbed A), but whether they do is a property of the condition that only labels can certify. Section 5.2 shows such a rule failing once the incumbent changes.

4 CERTIFIED POLICY SELECTION

Proposition 4 rules out label-free guarantees, and Proposition 2 shows that the benefit a gate can add is often small. Both point to the same design: choose among a few frozen candidates using target labels, and replace the incumbent only when the calibration sample *certifies* an improvement. Algorithm 1 implements this. It adapts high-confidence policy improvement (Thomas et al., 2015) to the structure of test-time adaptive re-ranking. Gate thresholds come from a grid $\mathcal{T}$ of quantiles of z on *unlabeled* traffic, so the gate class is fixed before any label is seen. Gates are fitted on one half of the sample and certified on the other; fixed actions need no fitting and are certified on the whole sample.

Lower-bound routines. $\mathrm{LB}_{\mathrm{bet}}$ is the predictable plug-in empirical-Bernstein bound of Waudby-Smith & Ramdas (2024). It is valid at every sample size for differences in $[-1, 1]$ and has a time-uniform variant. $\mathrm{LB}_{\mathrm{MP}}$ is the empirical-Bernstein bound of Maurer & Pontil (2009), which gives an explicit radius. $\mathrm{LB}_{\mathrm{norm}}$ is the normal approximation $\bar{D} - \zeta_{\delta/K}\hat{\sigma}/\sqrt{|I_c|}$ with $\zeta_\alpha = \Phi^{-1}(1-\alpha)$; it is valid only asymptotically.

Theorem 1 (Safety and regret of CPS)**.** *Assume the calibration queries are i.i.d. from the deployment distribution, utilities lie in $[0, 1]$, and $\mathcal{T}$ is fixed independently of the labels. For a Route candidate, θ_c is the value gap of its* fitted *gate. Suppose that each lower bound satisfies $\Pr(\ell_c \le \theta_c) \ge 1 - \delta/K$, conditionally on I_{fit} for Route candidates. Let $c^\star = \arg\max_{c \in \mathcal{C}} \theta_c$.*

(i) Safety. *With probability at least $1 - \delta$, CPS either keeps b or deploys an action whose value exceeds $V(b) + \epsilon$.*

(ii) Oracle inequality. *On the same event, the regret of the returned action relative to $c^\star$ is at most $\epsilon + (\theta_{c^\star} - \ell_{c^\star})$, and it is zero whenever b is optimal.*

(iii) Rate. *With $\mathrm{LB}_{\mathrm{MP}}$, let $\Lambda = \ln(2K/\delta)$ and let σ_c be the standard deviation of $U_c - U_b$. With probability at least $1 - 3\delta$, the regret relative to $c^\star$ is at most*

$$\epsilon + 2\sigma_{c^\star}\sqrt{2\Lambda/n_{c^\star}} + C\,\Lambda/(n_{c^\star} - 1),$$

where $n_c = |I_c|$ and $C \le 18$ is an absolute constant.

$\mathrm{LB}_{\mathrm{bet}}$ and $\mathrm{LB}_{\mathrm{MP}}$ satisfy the hypothesis in finite samples. $\mathrm{LB}_{\mathrm{norm}}$ satisfies it as $n \to \infty$ for fixed K by the central limit theorem, so (i) and (ii) hold asymptotically; Section 5.3 measures its finite-n coverage on the stored data. The loss from fitting a gate on half the sample is controlled separately:

Lemma 1 (Gate fitting)**.** *Let $\mathcal{G}$ be the $2|\mathcal{T}|$ threshold gates of Algorithm 1 for one route pair, and let $\hat{g}$ maximize the empirical gain on $n_f = |I_{\mathrm{fit}}|$ queries. With probability at least $1 - \delta$, $V(\textsc{Route}_{\hat{g}}) \ge \max_{g \in \mathcal{G}} V(\textsc{Route}_g) - 2\sqrt{2\ln(4|\mathcal{T}|/\delta)/n_f}$.*

With a union bound over the J route pairs, the regret relative to the best action in $\mathcal{A} \cup \{\textsc{Route}_g : g \in \mathcal{G}\}$ is at most $\epsilon + O\big(\sigma\sqrt{\log(K/\delta)/n} + \sqrt{\log(J|\mathcal{T}|/\delta)/n}\big)$ with probability $1 - 4\delta$. The second term, the price of fitting a gate, is conservative (0.33 at $n = 1{,}000$), but it grows only logarithmically in the grid size; fixed candidates pay only the first.

Corollary 1 (How many labels)**.** *Let $\theta_c > \epsilon$. Under the normal approximation, candidate c is certified with probability $1/2$ once*

$$n_{50}(c) = \Big(\frac{\zeta_{\delta/K}\,\sigma_c}{\theta_c - \epsilon}\Big)^2 \tag{2}$$

paired differences are available, and with probability at least $1 - \delta/K$ once $4n_{50}(c)$ are. For the betting bound the leading term is $2\ln(K/\delta)\,\sigma_c^2/(\theta_c - \epsilon)^2$. Route candidates are certified on half the sample, which doubles their requirement.

Equation (2) is the standard sample-size formula for a one-sided paired test (Webber et al., 2008; Sakai, 2016). It replaces fixed rules of thumb for the calibration size: the budget depends on the signal-to-noise ratio θ_c/σ_c, which spans two orders of magnitude in our data. It needs θ_c and σ_c, which a small pilot estimates poorly, so we recommend sequential labeling over a pilot-based plan.

Corollary 2 (Resolving the value of routing)**.** *Let a^+ be the better of s and a, and σ_R the standard deviation of $U_{\textsc{Route}_g} - U_{a^+}$. Under the normal approximation, the paired level-δ test of $V(\textsc{Route}_g) > V(a^+)$ reaches power $1/2$ only when $n \ge (\zeta_\delta \sigma_R/\Gamma(g))^2$, and $\Gamma(g) \le H_z$ by Proposition 2.*

CPS need not resolve this comparison to be safe, since it certifies every candidate against the incumbent. But *knowing* that routing beats the better simple action takes 220 to 790 labeled queries on our Quora conditions ($\Gamma \le 0.03$, $\sigma_R \approx 0.13$–0.33), and thousands where $\Gamma \approx 0.001$.

Sequential labeling. In production, labels arrive in batches, and recomputing a fixed-n bound after every batch invalidates it.

Corollary 3 (Anytime-valid CPS)**.** *Fix the candidates, including gates fitted on a separate pilot sample, and let labeled queries arrive i.i.d. For each candidate let $(\ell_{c,t})_{t \ge 1}$ be a lower confidence*

sequence at level δ/K, i.e. $\Pr(\exists t : \ell_{c,t} > \theta_c) \leq \delta/K$. *Then for any stopping time τ, possibly chosen by looking at the data, CPS applied at time τ deploys an action with $\theta_c \leq \epsilon$ with probability at most δ.*

The betting bound has such a variant (Howard et al., 2021; Waudby-Smith & Ramdas, 2024). The procedure is: fix $\mathcal{A}$, b, the utility, δ, ϵ and the grid; fit gates on a pilot; label queries in random order; deploy the first candidate whose bound exceeds ϵ, or keep b when the budget is exhausted; re-audit when candidates, traffic or utility change.

5 Experiments

5.1 Replay protocol

Every action is deterministic given the query and its pool, so we evaluate selection procedures by *replay*: from the stored per-query outputs of a condition with N judged queries, we draw $n \leq N/2$ calibration queries without replacement, run the procedure, and score the chosen action on all N queries. A deployment is *harmful* if the chosen action is worse than the incumbent on the full condition. Regret is measured against the best action, including the population-best threshold gate for each Route pair. Each cell is repeated 400 times (200 for the extended Quora action set and the δ and margin sweeps). We use $\delta = 0.1$ and $\epsilon = 0$. Route gates threshold pre-update score entropy on 19 quantiles of its unlabeled distribution. Appendix B lists every configuration.

Testbed A (audit): 23 conditions of a fixed-configuration protocol, in which the top-50 BGE-small (Xiao et al., 2023) candidates of seven BEIR datasets (Thakur et al., 2021) are re-ranked by a frozen scorer, BERT-base, BERT-large, DistilBERT (Devlin et al., 2019; Sanh et al., 2019) or, in six conditions ("bge*"), the retriever's own encoder, each with one adaptive configuration (three anchors, three steps, step size 0.3, with two exceptions). Aggregates use the 19 conditions with $N \geq 300$. **Qilin:** a search log (Chen et al., 2025) with 6,192 test queries, platform result lists and click labels; the static action is a trained cross-encoder, and the adaptive actions adapt a dense query embedding with one, three or five anchors. **Testbed B (self-adaptation):** the retriever's encoder adapts its own query embedding on ten BEIR datasets, over anchor depth $\{1, 2, 3\}$, step size $\{0.3, 0.1, 0.03, 0.01\}$ and steps $\{1, 3\}$ (24 adaptive actions).

5.2 Audit: recommendations depend on the action set

Table 1 compares fixed-configuration recommendations with complete action sets. Of the 22 conditions with a Skip or All recommendation, 15 lose to the retriever's order, and the mean difference over all 22 is -0.085 MRR@10. The remaining seven are the six retriever-encoder conditions, where Skip coincides with the retriever's order, and TREC-COVID with BERT-base (50 queries), where All beats it by 0.019. Tuning step size and anchor depth raises All wherever a grid exists and reverses the Quora recommendation, yet tuned adaptation stays below the retriever's order in every row. On Qilin, one anchor instead of three raises All from 0.578 to 0.710, well above the three-anchor Route (0.588); the platform's order is higher still (0.736), though clicks collected on it favor it (Joachims et al., 2017).

CPS recovers these reversals from labels: with the full action sets (36 candidates on Quora), it replaces Skip by an adaptive action in 70% (Quora) and 97% (Qilin) of calibrations at $n = 100$, and at the system level it keeps the retriever's or platform's order in every calibration with $n \geq 50$ (Appendix C).

Label-free rules and frozen decisions. The rule "adapt unless score entropy is high" works well within Testbed A: its best threshold (2.0, chosen in-sample) picks the better of Skip and All in 17 of 19 conditions (mean regret 0.0007). But it has no notion of the incumbent: as a deployment rule it replaces the retriever's order with a worse ranking in 14 of 19 conditions (mean -0.121). Decisions also fail to transfer: the best scorer-level action of one condition, deployed on another condition of the same dataset with a different scorer, is wrong in 30 of 56 pairs (mean regret 0.029, maximum 0.189). Proposition 4 allows both failures: nothing label-free marks the conditions in which the rule is wrong.

Table 1: Fixed-configuration recommendations versus complete action sets (MRR@10). "Best with tuned adaptation": best existing run over step sizes, steps and anchor depths; "–": no grid was run.

Condition	Fixed-configuration choice	Best with tuned adaptation	Retriever order
Quora / BERT-large	SKIP 0.587	ALL, 2 anchors, step 0.075 0.685	**0.880**
NQ / BERT-large	ALL 0.241	ALL, 1 anchor, step 0.2 0.297	**0.386**
HotpotQA / BERT-base	ALL 0.609	–	**0.834**
SciDocs / BERT-large	ALL 0.253	ALL, 1 anchor, step 0.2 0.264	**0.334**
NFCorpus / BERT-large	ALL 0.448	ALL, 2 anchors, step 0.3 0.490	**0.530**
Qilin search log / CE	ROUTE 0.588	ALL, 1 anchor 0.710	**0.736**

Table 2: Harmful-deployment rate (%) and regret ($\times 10^{-3}$ MRR@10) at $n = 50$ labeled queries, averaged over conditions (Testbed A: 19; Testbed B: 9; Qilin: one condition with 1, 3 and 5 anchors, platform order as incumbent). Upper block: uncertified or uncorrected selectors; lower block: CPS with three lower-bound routines. $\delta = 0.1$.

	A / scorer		A / system		B / system		Qilin / system	
Method	harm	regret	harm	regret	harm	regret	harm	regret
uncertified best	22.9	9.1	53.8	9.6	70.2	9.1	64.0	15.7
uncertified, margin 0.01	14.6	8.6	27.9	6.0	51.3	7.8	24.5	10.0
paired t-test, no correction	5.7	9.9	9.6	1.9	28.3	5.8	10.2	2.1
select-then-certify (normal)	0.9	27.7	1.5	0.5	2.4	6.0	2.0	0.5
CPS, normal bound	0.7	13.3	0.2	0.3	0.1	5.7	0.0	0.0
CPS, Maurer–Pontil	0.0	85.2	0.0	0.2	0.0	7.1	0.0	0.0
CPS, betting bound	0.0	70.8	0.0	0.2	0.0	7.1	0.0	0.0
always keep the incumbent	0.0	85.2	0.0	0.2	0.0	7.1	0.0	0.0

5.3 CERTIFICATION CONTROLS HARM; ITS REGRET COST SHRINKS WITH LABELS

Table 2 and Figure 1 compare CPS with other selectors. The uncertified empirical best, the natural "direct evaluation" baseline, deploys an action worse than the incumbent in 23–70% of calibrations at $n = 50$. This is not a near-tie artifact: counting only deployments worse by more than 0.005 MRR@10 still gives 14% and 42% on Testbed A, 44% on Testbed B and 51% on Qilin. An uncorrected paired t-test lowers harm to 6–28%, but ignores selection over K candidates; with $K = 48$ on Testbed B the winner's curse keeps it at 28%. CPS with the normal bound keeps harm at or below 0.7% on average for every n and utility (worst single condition 5%); pooled over all scorer-level calibrations, 1.4% of its deployments are harmful, against 25% for the uncertified choice. The finite-sample bounds never deploy a harmful action, but at $n = 50$ they rarely deploy at all, and their regret approaches that of keeping the incumbent (last row of Table 2). Certifying only the half-sample winner avoids the multiplicity correction, but selects on half the data; it has higher harm and regret than CPS in every column. The normal bound is only asymptotically valid, so we checked it at the Bonferroni tails CPS uses, over 84 real action pairs ($N \geq 400$): its average miscoverage matches the level at $\alpha \in \{0.1, 0.05\}$ (ratio 0.99, 1.02 at $n = 50$), exceeds it at deep tails (ratio 1.7 at $\alpha = 0.002$, i.e. $K = 48$, worst pair 1.2%), and is conservative by $n = 200$ (ratio ≤ 0.89); the betting bound never exceeds its level. Our recommended operating point is therefore CPS with the normal bound for one-shot audits with few candidates or at least 100 queries, and the betting confidence sequence when K is large, labels arrive sequentially, or a finite-sample guarantee is required.

Safety costs regret when the incumbent is weak and labels are scarce: at the scorer level and $n = 50$, CPS defers in half of the calibrations and its regret is 0.013, against 0.009 uncertified; the gap is 0.002 by $n = 200$. A tuned minimum-gain margin can match CPS on Testbed A (0.04 gives 1.8% harm at regret 0.009; Figure 3), but its harm is not calibrated: the same margin gives 5.2% on Testbed B (worst dataset 14%), and 0.02 gives 10.9% at the system level of Testbed A. CPS controls harm with one parameter, δ, across testbeds.

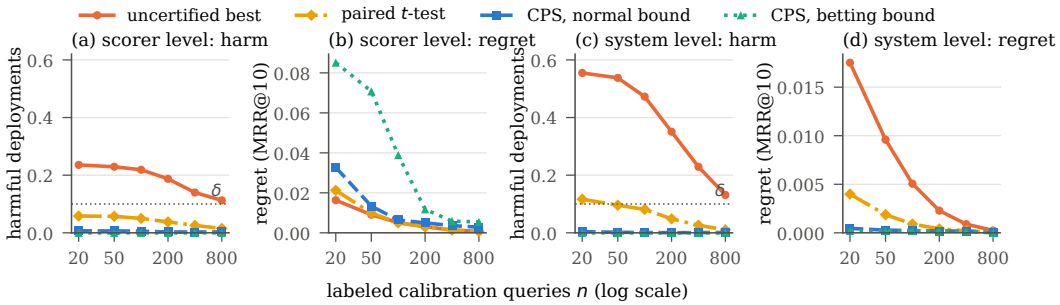

Figure 1: Harm and regret versus the number of labeled queries on Testbed A (MRR@10; 19 conditions for $n \leq 100$, 15 for $n \in \{200, 400\}$, 11 for $n = 800$, since $n \leq N/2$). Incumbent: the frozen scorer (scorer level) or the retriever's order (system level). Dotted: δ.

5.4 The label budget is predictable given the effect size

For every ordered pair of actions with a positive gap, we compare the median certification point predicted by Corollary 1 (population θ and σ, one candidate) with the empirical one from 22 sample sizes $\times$ 300 draws (Figure 2a). For the normal bound, the median error is 6% and the Spearman correlation 0.98 over the 33 pairs whose certification point lies in the grid (7.5% and 0.94 for the 20 pairs with $N \geq 300$ and certification below $N/10$). Budgets range from 8 to over 3,000 labels. For the betting bound, the leading term is too small by a stable factor: fitted on half of the conditions, it (2.7–2.9) predicts the other half within a factor two for 95% of pairs.

A pilot estimates the effect size poorly: plugging pilot estimates into Equation (2) lands within a factor two for only 42% of the 27 pairs with $N \geq 300$ with 30 pilot queries, and 59% with 100. Sequential labeling avoids this if the bound is time-uniform. For the 64 pairs whose challenger is strictly worse, recomputing the normal bound after every label and deploying at the first certification selects the harmful challenger in 6.3% of label streams (worst pair 42%); the time-uniform betting sequence in 0.02% (worst 1%).

5.5 Why routing rarely pays

Figure 2b measures the rungs of Proposition 2 for Route(Skip,All). The per-query headroom $H = \min(G, L)$ lies between 0.02 and 0.18 MRR@10, so in every condition some queries are helped and some are hurt. We bracket H_z for score entropy between a two-fold cross-fitted binned estimate, which is biased downward, and an in-sample one, which is biased upward. Both are below 0.01 in every condition except the highest-entropy Quora slice (0.010 to 0.028, against $H = 0.176$). The cross-fitted estimate is positive only in three Quora conditions (0.006–0.010, at most 15% of H); the in-sample estimate reaches at most 21.5% of H, in a 323-query condition where the cross-fitted estimate is negative. The pre-update anchor loss sees more on the Quora slice (0.074) and little elsewhere. A population-fitted entropy gate beats the better simple action by more than 0.005 in only these three Quora conditions, and by Corollary 2 knowing that it does takes 220–790 labeled queries.

Anchor regimes. Stratifying each condition by entropy quintile and estimating γ, λ with labels gives 52 strata in regime I, 28 in II, 14 in III and one in IV (of 95). In the 17 conditions whose majority regime is I or II, it predicts the better of Skip and All in all 17; the two majority-III conditions are the Quora slice, where Route wins by 0.029, and NQ with the retriever's encoder, where $\rho < 0.06$ stays below $\rho^\star$ (details in Appendix E). In 9 of 19 conditions a single regime covers all entropy strata, so no gate on those strata can pay.

Mechanism: adaptation pulls the scorer toward the retriever. The anchors are the retriever's top candidates, which explains regime I. Per query, the gain of All over Skip correlates with the retriever's advantage over the frozen scorer ($r = 0.24$–0.68, median 0.56, over the 14 conditions with $N \geq 300$ and a distinct scorer; 0.61 and 0.94 on Qilin with three and one anchors). Where the retriever beats the scorer, adaptation helps on 28–73% of queries; where it is worse, adaptation helps on only 4–20% and hurts on 38–93%. Hence adaptation never overtakes a much stronger retriever (Table 1), and a gate must predict per query whether the retriever or the scorer is right, which the

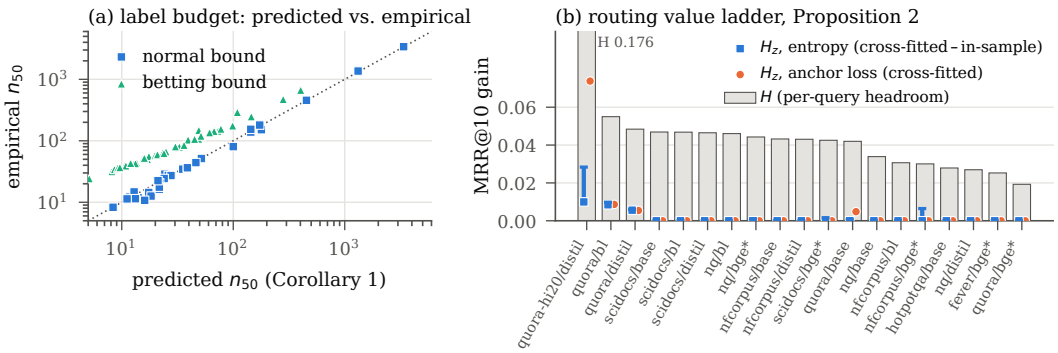

Figure 2: (a) Predicted vs. empirical median certification point per action pair with a positive gap; dotted: $y = x$. (b) Per condition ($N \geq 300$): headroom H (bars), H_z for entropy between cross-fitted and in-sample estimates (blue), and cross-fitted H_z for the anchor loss (orange).

scorer's score shape barely reveals. The anchor–scorer overlap in the top 10, a signal suggested by this mechanism, raises the cross-fitted H_z on Quora (to 0.015 and 0.009) but is zero elsewhere: it shows where the rankings differ, not which is right. Our conclusions about adaptation therefore concern re-rankers weaker than their first stage.

5.6 Self-adaptation of the retriever's encoder

In Testbed B the retriever adapts its own queries, so there is no weaker scorer to pull toward the retriever, and adaptation helps only on ArguAna (17 of 24 configurations, best +0.040 MRR@10; Appendix Table 5). Elsewhere, population-fitted gates gain at most 0.011. With $K = 48$ candidates the uncertified choice deploys a harmful action in 70% of calibrations at $n = 50$ and the t-test in 28%, against 0.1% for CPS with the normal bound, which on ArguAna deploys the improvement in 82% of calibrations at $n = 100$. The finite-sample bounds never certify the ArguAna gain up to $n = 400$ at this K.

6 Related Work

Selective TTA methods such as EATA and SAR update only on reliable samples (Niu et al., 2022; 2023), and TTA gains depend on hyperparameters that need target labels to tune (Zhao et al., 2023); dense PRF helps on some collections and hurts on others (Yu et al., 2021; Li et al., 2023). A Route gate is a treatment rule: Proposition 2 is the value of the first-best rule (Manski, 2004), and empirical welfare maximization bounds estimated rules (Kitagawa & Tetenov, 2018; Athey & Wager, 2021). CPS follows high-confidence and safe policy improvement (Thomas et al., 2015; Laroche et al., 2019) with empirical-Bernstein and betting bounds (Maurer & Pontil, 2009; Waudby-Smith & Ramdas, 2024); the label budget parallels power analysis in IR evaluation (Webber et al., 2008; Sakai, 2016). These tools are not new; our contribution is to cast TTA deployment in these terms and to bound and measure what gating and uncertified selection cost (more in Appendix F).

7 Limitations and Conclusion

Limitations. The guarantees assume exchangeable calibration and deployment traffic, and we measure drift only through transfer between conditions. Each benchmark is a single split, all adaptive actions follow one recipe, every re-ranker is weaker than its first stage, and Qilin's clicks favor the displayed order; the theory and CPS do not depend on this, the empirical findings do. The normal bound is asymptotic and slightly anti-conservative at deep tails with few labels, the finite-sample bounds rarely deploy below a few hundred labels, and Lemma 1 is loose.

Conclusion. Whether to adapt a re-ranker at test time depends on the action set, the incumbent and the utility, which label-free statistics cannot settle; certified policy selection makes it a labeled audit with explicit risk and a predictable label budget.

Reproducibility Statement

All results are computed from stored per-query outputs by scripts in the supplement: condition loaders, the CPS implementation with all three bounds, replay drivers for every table and figure, and the Testbed B launch and scheduling scripts. Every number in the paper is rendered from these artifacts. Appendix B lists datasets, query counts, model configurations and the exact adaptation recipe, and Appendix A contains complete proofs with their assumptions.

AI Use Statement

Generative AI tools assisted with organizing and revising the manuscript, drafting and checking mathematical statements and proofs, writing analysis and plotting scripts, auditing the consistency between stored experimental artifacts and reported numbers, and preparing LaTeX. All experimental numbers are computed from stored per-query outputs by scripts that the authors reviewed. The authors take responsibility for the content.

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, November 2023. doi: 10.1126/science.adi6000.

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, June 2025. doi: 10.1214/24-AOAS1998. Preprint: arXiv:2110.01052.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, January 2021. doi: 10.3982/ECTA15732.

Benjamin A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems*, 30(1):1–34, February 2012. doi: 10.1145/2094072.2094076.

Jia Chen, Qian Dong, Haitao Li, Xiaohui He, Yan Gao, Shaosheng Cao, Yi Wu, Ping Yang, Chen Xu, Yao Hu, Qingyao Ai, and Yiqun Liu. Qilin: A multimodal information retrieval dataset with app-level user sessions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3670–3680. Association for Computing Machinery, 2025. doi: 10.1145/3726302.3730279.

Steve Cronen-Townsend, Yunhua Zhou, and W Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 299–306. Association for Computing Machinery, 2002. doi: 10.1145/564376.564429.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, Bellevue, Washington, USA, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017. URL https://arxiv.org/abs/1705.08500.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, April 2021. doi: 10.1214/20-AOS1991.

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, New York, NY, USA, 2017. Association for Computing Machinery. doi: 10.1145/3018661.3018699.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, March 2018. doi: 10.3982/ECTA13288.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *ACM Transactions on Information Systems*, 41(3):1–40, 2023. doi: 10.1145/3570724.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, July 2004. doi: 10.1111/j.1468-0262.2004.00530.x.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, Montreal, Quebec, Canada, June 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. *arXiv preprint arXiv:2006.01862*, 2020. URL https://arxiv.org/abs/2006.01862.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. *arXiv preprint arXiv:2204.02610*, 2022. URL https://arxiv.org/abs/2204.02610.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. URL https://arxiv.org/abs/2302.12400.

Joseph J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323. Prentice Hall, 1971.

Tetsuya Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 5–14, New York, NY, USA, 2016. Association for Computing Machinery. doi: 10.1145/2911451.2911492.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gilad Markovits. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems (TOIS)*, 30(2): 1–35, 2012. doi: 10.1145/2180868.2180873.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/thomas15.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. Pseudo-relevance feedback for multiple representation dense retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, pp. 297–306, New York, NY, USA, 2021b. Association for Computing Machinery. doi: 10.1145/3471158.3472250.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

William Webber, Alistair Moffat, and Justin Zobel. Statistical power in retrieval experimentation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pp. 571–580, New York, NY, USA, 2008. Association for Computing Machinery. doi: 10.1145/1458082.1458158.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023. URL https://arxiv.org/abs/2309.07597.

HongChien Yu, Chenyan Xiong, and Jamie Callan. Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pp. 3592–3596, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3459637.3482124.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642. Curran Associates, Inc., 2022. doi: 10.52202/068431-2799. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fc28053a08f59fccb48b11f2e31e81c7-Abstract-Conference.html.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

Yunhua Zhou and W Bruce Croft. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 543–550. Association for Computing Machinery, 2007. doi: 10.1145/1277741.1277835.

A PROOFS

Throughout, $U_a(q) \in [0,1]$, so every per-query difference lies in $[-1,1]$ and every expectation below is finite.

A.1 PROPOSITION 1

For any gate g, $U_{\text{ROUTE}_g} = U_s + g\Delta$. Taking expectations and splitting $\Delta = \Delta_+ - \Delta_-$ gives $V(\text{ROUTE}_g) - V(s) = \mathbb{E}[g\Delta_+] - \mathbb{E}[g\Delta_-] = G_g - L_g$. Since $V(a) - V(s) = G - L$, subtracting yields $V(\text{ROUTE}_g) - V(a) = (G_g - L_g) - (G - L) = (L - L_g) - (G - G_g)$. Both differences are positive exactly under the stated conditions. □

A.2 PROPOSITION 2

(ii). Let g be any measurable map of z. By the tower property, $V(\text{ROUTE}_g) - V(s) = \mathbb{E}[g(z)\Delta] = \mathbb{E}[g(z)\mu(z)]$. This is maximized pointwise by $g^\star(z) = \mathbb{1}[\mu(z) > 0]$, with value $\mathbb{E}[\mu_+]$. Likewise $V(\text{ROUTE}_g) - V(a) = \mathbb{E}[(g(z) - 1)\mu(z)]$ is maximized by the same $g^\star$, with value $\mathbb{E}[\mu_-]$. The two objectives differ by the constant $V(a) - V(s)$, so a single gate maximizes both, and

$$\sup_g \Big(V(\text{ROUTE}_g) - \max\{V(s), V(a)\} \Big)$$
$$= \min\Big(\sup_g \big(V(\text{ROUTE}_g) - V(s) \big),\ \sup_g \big(V(\text{ROUTE}_g) - V(a) \big) \Big) = \min(\mathbb{E}[\mu_+], \mathbb{E}[\mu_-]).$$

(i) For any gate $g \in \{0,1\}$, possibly depending on the labels, $\mathbb{E}[g\Delta] \le \mathbb{E}[\Delta_+] = G$ and $\mathbb{E}[(g-1)\Delta] \le \mathbb{E}[\Delta_-] = L$, so $\Gamma(g) \le \min(G, L)$; the gate $\mathbb{1}[\Delta > 0]$ attains both bounds, i.e. (ii) with $z = \Delta$. *(iii).* $x \mapsto x_+$ and $x \mapsto x_-$ are convex, so Jensen's inequality conditional on z gives $\mu(z)_+ \le \mathbb{E}[\Delta_+ \mid z]$ and $\mu(z)_- \le \mathbb{E}[\Delta_- \mid z]$. Taking expectations, $\mathbb{E}[\mu_+] \le G$ and $\mathbb{E}[\mu_-] \le L$, hence $H_z \le H$; nonnegativity holds because $\mu_\pm \ge 0$. If, conditionally on z, Δ lies almost surely on one side of zero, both Jensen inequalities are equalities, so $H_z = H$. Finally, $H_z = 0$ iff $\mathbb{E}[\mu_+] = 0$ or $\mathbb{E}[\mu_-] = 0$, i.e. iff $\mu \le 0$ a.s. or $\mu \ge 0$ a.s. □

A.3 PROPOSITION 3

By the law of total expectation conditional on z, $\mu(z) = \rho(z)\mathbb{E}[\Delta \mid z, Y = 1] + (1 - \rho(z))\mathbb{E}[\Delta \mid z, Y = 0] = \rho\gamma - (1-\rho)\lambda$. In regime I both terms are nonnegative and the first is positive whenever $\rho > 0$ (if $\rho = 0$ the regime is vacuous at z); regime II is symmetric. In regime III, $\mu > 0 \iff \rho(\gamma + \lambda) > \lambda \iff \rho > \rho^\star$. In regime IV write $\gamma = -|\gamma|$, $\lambda = -|\lambda|$; then $\mu = (1-\rho)|\lambda| - \rho|\gamma| > 0 \iff \rho < |\lambda|/(|\gamma| + |\lambda|) = \rho^\star$, and $\mu = 0$ if $\gamma = \lambda = 0$. The statement about H_z follows from Proposition 2(iii): $H_z > 0$ iff μ takes both signs with positive probability, and the sign of μ is determined by the regime and, in regimes III and IV, by the side of $\rho^\star$. □

A.4 PROPOSITION 4

Let Q be the set of queries on which the top-k lists of a_1 and a_2 differ; by assumption $\Pr(Q) > 0$. Outside Q the two lists coincide, so $U_{a_1} = U_{a_2}$ there under any labels. For $q \in Q$, let $j(q)$ be the first rank at which the lists differ, $d^{(1)}(q)$ the document that a_1 places at rank $j(q)$, and $d^{(2)}(q)$ the document that a_2 places there. These are distinct, and each appears at a rank $> j(q)$, or not at all, in the other list. Start from any condition c_0 and define c_1 and c_2 by keeping the query distribution, candidate pools and all model outputs of c_0, and changing only the relevance labels on Q: in c_1 the only relevant document of $q \in Q$ is $d^{(1)}(q)$, in c_2 it is $d^{(2)}(q)$. A label-free observation O is a function of queries, pools and model outputs, so its law is the same in c_1 and c_2. For reciprocal rank at k, a_1 scores $1/j(q)$ and a_2 strictly less on every $q \in Q$ in c_1, and conversely in c_2. The same holds for average precision at k. Hence $d_1 = V_{c_1}(a_1) - V_{c_1}(a_2) > 0$ and $d_2 = V_{c_2}(a_2) - V_{c_2}(a_1) > 0$. Let $p = \Pr(\psi(O) = a_1)$, which is the same in both conditions. The regrets are $(1-p)d_1$ in c_1 and $p\,d_2$ in c_2. Their maximum is minimized where $(1-p)d_1 = p\,d_2$, i.e. at $p = d_1/(d_1 + d_2)$, where it equals $d_1 d_2/(d_1 + d_2) \ge \min(d_1, d_2)/2$. □

A.5 THEOREM 1

Conditional on I_{fit}, each Route candidate is a fixed action, and its differences on I_{cert} are i.i.d. with mean θ_c because I_{cert} is independent of I_{fit}. Fixed actions have i.i.d. differences on the whole sample. Let $E_1 = \bigcap_{c\in\mathcal{C}}\{\ell_c \le \theta_c\}$. For Route candidates, $\Pr(\ell_c > \theta_c) = \mathbb{E}[\Pr(\ell_c > \theta_c \mid I_{\mathrm{fit}})] \le \delta/K$; for fixed candidates the bound holds directly. By the union bound, $\Pr(E_1) \ge 1-\delta$. *(i).* On E_1, if CPS deploys $\hat{c} \ne b$, then $\theta_{\hat{c}} \ge \ell_{\hat{c}} > \epsilon$. *(ii).* Work on E_1. If b is optimal, i.e. $\theta_c \le 0$ for all c, then $\ell_c \le \theta_c \le 0 \le \epsilon$ for every c, so CPS keeps b and its regret is zero. Otherwise, if CPS deploys $\hat{c}$, then $\theta_{\hat{c}} \ge \ell_{\hat{c}} \ge \ell_{c^\star}$, so the regret $\theta_{c^\star} - \theta_{\hat{c}} \le \theta_{c^\star} - \ell_{c^\star}$. If CPS keeps b, then $\ell_{c^\star} \le \epsilon$, and the regret is $\theta_{c^\star} \le \epsilon + (\theta_{c^\star} - \ell_{c^\star})$. *(iii).* For $X \in [0,1]$ i.i.d., Theorem 4 of Maurer & Pontil (2009) gives, with probability at least $1-\alpha$, $\mathbb{E}X \le \bar{X} + \sqrt{2\hat{V}\ln(2/\alpha)/n} + 7\ln(2/\alpha)/(3(n-1))$, with $\hat{V}$ the unbiased sample variance. Applying it to $(D+1)/2$ and to $(1-D)/2$ for $D = U_c - U_b \in [-1,1]$ gives the two-sided statement $|\bar{D}_c - \theta_c| \le r_c := \sqrt{2\hat{V}_c\Lambda/n_c} + 14\Lambda/(3(n_c-1))$, each side with probability at least $1-\delta/K$, where $\Lambda = \ln(2K/\delta)$. $\mathrm{LB}_{\mathrm{MP}}$ is $\ell_c = \bar{D}_c - r_c$. Let $E_2 = \bigcap_c\{\theta_c \le \bar{D}_c + r_c\}$, so that $\Pr(E_2) \ge 1-\delta$. By Theorem 10 of the same paper, applied to $(D+1)/2$ at level δ/K, with probability at least $1-\delta/K$ we have $\sqrt{\hat{V}_c} \le \sigma_c + 2\sqrt{2\ln(K/\delta)/(n_c-1)}$. Call the intersection over c of these events E_3; then $\Pr(E_3) \ge 1-\delta$. On $E_1 \cap E_2$, part (ii) gives regret at most $\epsilon + \theta_{c^\star} - \ell_{c^\star} \le \epsilon + 2r_{c^\star}$. On E_3, using $\ln(K/\delta) \le \Lambda$ and $\sqrt{n_c(n_c-1)} \ge n_c - 1$,

$$2r_c \le 2\sigma_c\sqrt{2\Lambda/n_c} + 8\Lambda/(n_c-1) + 28\Lambda/(3(n_c-1)) \le 2\sigma_c\sqrt{2\Lambda/n_c} + 18\Lambda/(n_c-1).$$

The probability of $E_1 \cap E_2 \cap E_3$ is at least $1-3\delta$. □

A.6 LEMMA 1

For a fixed gate g, the variables $g(z(q_i))\Delta(q_i)$ lie in $[-1,1]$ and have mean $V(\textsc{Route}_g) - V(s)$. By Hoeffding's inequality, their empirical mean over n_f queries deviates from this mean by more than t with probability at most $2\exp(-n_f t^2/2)$. The grid is fixed before labels are seen, so a union bound over the $2|\mathcal{T}|$ gates shows that, with probability at least $1-\delta$ and $t = \sqrt{2\ln(4|\mathcal{T}|/\delta)/n_f}$, all empirical gains are within t of their means. On this event, for the maximizer $\hat{g}$ and any $g \in \mathcal{G}$, $V(\textsc{Route}_{\hat{g}}) - V(s) \ge \widehat{\mathrm{gain}}(\hat{g}) - t \ge \widehat{\mathrm{gain}}(g) - t \ge V(\textsc{Route}_g) - V(s) - 2t$. □

The end-to-end statement after Lemma 1 follows by applying the lemma to each of the J route pairs at level δ/J, so that every fitted Route candidate is within $2t$ of the best grid gate of its pair, and combining with Theorem 1(iii) on the event $E_1 \cap E_2 \cap E_3$. The best fitted candidate is then within $2t$ of the best action in $\mathcal{A} \cup \{\textsc{Route}_g\}$.

A.7 COROLLARIES

Corollary 1. Under the normal approximation, $\ell_c = \bar{D}_c - \zeta_{\delta/K}\hat{\sigma}_c/\sqrt{n}$ with $\bar{D}_c \approx \mathcal{N}(\theta_c, \sigma_c^2/n)$ and $\hat{\sigma}_c \approx \sigma_c$. Then $\Pr(\ell_c > \epsilon) \approx \Phi\big(\sqrt{n}(\theta_c - \epsilon)/\sigma_c - \zeta_{\delta/K}\big)$. This equals $1/2$ at $n = n_{50}$, and $1-\delta/K$ at $\sqrt{n}(\theta_c-\epsilon)/\sigma_c = 2\zeta_{\delta/K}$, i.e. at $n = 4n_{50}$. For the betting bound, the leading term of the half-width at fixed n is $\sqrt{2\sigma^2\ln(K/\delta)/n}$ (Waudby-Smith & Ramdas, 2024), which gives the stated leading term. *Corollary 2.* The paired normal test rejects when $\bar{D}_R > \zeta_\delta\hat{\sigma}_R/\sqrt{n}$, with $D_R = U_{\textsc{Route}_g} - U_{a^+}$ and mean $\Gamma(g)$. Its power is approximately $\Phi(\sqrt{n}\,\Gamma(g)/\sigma_R - \zeta_\delta)$, which is at least $1/2$ iff $n \ge (\zeta_\delta\sigma_R/\Gamma(g))^2$. Proposition 2(ii) gives $\Gamma(g) \le H_z$. □

A.8 COROLLARY 3

Let $E = \bigcap_{c\in\mathcal{C}}\{\ell_{c,t} \le \theta_c \text{ for all } t\}$. By the definition of a confidence sequence and the union bound, $\Pr(E) \ge 1-\delta$. On E, $\ell_{c,\tau} \le \theta_c$ for every c and every value of τ, whatever the stopping rule; hence if CPS deploys $\hat{c}$ at time τ, $\theta_{\hat{c}} \ge \ell_{\hat{c},\tau} > \epsilon$. Freezing the gates on a pilot makes every candidate a fixed action, so the differences seen after the pilot are i.i.d. with mean θ_c. □

B EXPERIMENTAL DETAILS

Notation. Table 3 collects the symbols used in the main text.

Table 3: Notation.

Symbol	Meaning
q, P, $\mathcal{D}_q$	query, query distribution of a condition, candidate pool of q
$\mathcal{A}$, b	finite action set, incumbent action ($b \in \mathcal{A}$)
$U_a(q) \in [0,1]$, $V(a) = \mathbb{E}U_a$	per-query utility and value of action a
$\Delta = U_a - U_s$, $\Delta_\pm$	per-query difference of two actions, its positive and negative parts
$G = \mathbb{E}\Delta_+$, $L = \mathbb{E}\Delta_-$	total gain and harm of a relative to s
$z(q)$, $g(z) \in \{0,1\}$	label-free signal, gate
$\mu(z) = \mathbb{E}[\Delta \mid z]$	conditional effect of adaptation given the signal
$H = \min(G, L)$, H_z	per-query and z-measurable routing headroom (Prop. 2)
Y, $\rho(z)$, $\gamma(z)$, $\lambda(z)$	informative-anchor indicator, its probability, conditional effect, loss
n, δ, ϵ	calibration size, risk level, minimum improvement
K, $\mathcal{T}$, J	number of candidates, threshold grid, number of route pairs
$\theta_c = V(c) - V(b)$	gap of candidate c to the incumbent
ζ_α	standard normal $(1-\alpha)$-quantile
ℓ_c, r_c, σ_c	lower bound, radius, s.d. of the paired difference for candidate c
ω, κ, τ, f_j	query offset, regularization weight, temperature, score (Eq. 1)
$\phi = (m, T, \eta)$	adaptation hyperparameters: anchor depth, steps, step size

Pipeline. All BEIR conditions use the top-50 candidates of BGE-small-en-v1.5 (Xiao et al., 2023), retrieved from the full corpus with stored embeddings; the retriever's order is the candidates sorted by retriever score. Only queries with at least one judged relevant document are scored. This gives 10,000 queries for Quora, 7,405 for HotpotQA, 6,666 for FEVER (judged subset), 3,452 for NQ, 1,000 for SciDocs, 323 for NFCorpus, 50 for TREC-COVID, 1,401 for ArguAna, 648 for FiQA and 300 for SciFact. The frozen scorer embeds the query and candidates with the stated encoder (mean-pooled final hidden states for the BERT scorers, stored embeddings for BGE-small), maps each embedding x to $\tanh(x/5)$ and then to the Poincaré ball of curvature 1 by the exponential map at the origin; scores are negative hyperbolic distances. The offset ω is added to the query's tangent vector before the exponential map. The loss is Equation (1); no model weights change. The implementation also has entailment-cone terms, which the code enables only for other objectives and which are therefore inactive in all runs. PRF anchors are the retriever's top m candidates. The value-ladder estimates of Section 5.5 use ten quantile bins of z.

Testbed A. The 23 conditions are NFCorpus, SciDocs, TREC-COVID, NQ and Quora, each with four scorers (BERT-base-uncased, BERT-large-uncased, DistilBERT-base-uncased, and the retriever's encoder); HotpotQA with BERT-base-uncased; FEVER; and a Quora slice consisting of the 20% of queries with the highest DistilBERT score entropy. The adaptive configuration is fixed at $\phi = (3, 3, 0.3)$, except for FEVER, $\phi = (3, 2, 0.3)$, and the Quora slice, $\phi = (1, 1, 0.4)$. Partial grids over $\eta \in \{0.15, 0.2, 0.3, 0.4\}$, $m \in \{1, 2, 3\}$ and $T \in \{1, 2, 3\}$ exist for BERT-large on Quora, NQ, SciDocs and NFCorpus (Table 1). In the six retriever-encoder conditions (five datasets and FEVER), the frozen scorer reproduces the retriever's order on 99.7–100% of queries. We keep queries on which both actions score zero.

Qilin. The Qilin search test split (Chen et al., 2025) has 6,192 queries with platform result lists (20.7 candidates on average, the first m of which are the anchors) and click labels. The static action is a cross-encoder trained on the Qilin training split. The adaptive actions use the dataset's precomputed dense query and note embeddings with the recipe above, $m \in \{1, 3, 5\}$, three steps and step size 0.3. These runs also undo an update whose anchor loss increases (70 of 6,192 queries with three anchors, two with one anchor); the undo is part of the action. The Route action of the fixed three-anchor configuration (0.588) uses a four-feature logistic gate fitted on 4,413 development queries; we re-derived all its test decisions from the stored model.

Testbed B. We use the same BEIR pools and the retriever's encoder as scorer, with the $3 \times 2 \times 4$ grid $\phi = (m, T, \eta) \in \{1, 2, 3\} \times \{1, 3\} \times \{0.3, 0.1, 0.03, 0.01\}$ and a steps-0 run, which reproduces the retriever's order on 99.7–100% of queries.

Replay and statistics. In each cell we draw 400 calibration samples without replacement (200 on the extended Quora action set and in the δ and margin sweeps; for Section 5.4, 300 per sample size on a grid of 22 sizes from 8 to 4,000, capped at N). Route gates choose a threshold among the 5%, ..., 95% quantiles of the signal over all queries of the condition, which uses no labels, together with a direction. CPS uses Bonferroni over the K non-incumbent candidates. The betting bound is the predictable plug-in empirical-Bernstein bound of Waudby-Smith & Ramdas (2024, Thm. 2) with truncation $c = 1/2$, applied to $(D+1)/2$. We checked coverage on a simulated heavy-zero difference distribution (70% zeros, 4,000 replicates; script in the supplement): at $n \in \{20, 50, 100, 400\}$ and nominal $\delta = 0.1$, the betting bound miscovers in at most 0.5% of replicates and the normal bound in 10–11%. All code and per-query artifacts are in the supplement.

C Extended Action Sets

Quora. The Quora/BERT-large action set contains the retriever's order, Skip, 18 All configurations (a partial grid over $\eta \in \{0.15, 0.2, 0.3, 0.4\}$, $m \in \{1, 2, 3\}$ and $T \in \{1, 2, 3\}$, two small-step runs with $\eta = 0.075$, and a run with BM25 anchors), and entropy-gated Route actions over Skip (and, at the system level, over the retriever's order): 36 non-incumbent candidates at the scorer level and 55 at the system level. At the scorer level, CPS with the normal bound replaces Skip in 28%, 70%, 94% and 100% of calibrations at $n = 50, 100, 200, 400$ (regret 0.076, 0.036, 0.013, 0.005) and never deploys a harmful action. The betting bound certifies in at most 1% of calibrations below $n = 400$. At the system level every CPS variant keeps the retriever's order at every n, whereas the uncertified choice deploys a worse action in 38% of calibrations at $n = 50$.

Qilin. The action set contains the platform order, the cross-encoder (Skip), All with 1, 3 and 5 anchors, and the corresponding Route actions. At the scorer level, CPS with the normal bound replaces Skip in 74% and 97% of calibrations at $n = 50$ and $n = 100$ (regret 0.048 and 0.009). At the system level CPS keeps the platform order for $n \geq 50$, and the uncertified choice deploys a worse action in 64% at $n = 50$ and 30% at $n = 200$.

D Additional Results

Harm–regret trade-off. Figure 3 sweeps $\delta \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$ for the certified selectors and the minimum-gain margin $\{0, 0.005, 0.01, 0.02, 0.04\}$ for the uncertified one, at $n = 50$ on Testbed A. At the system level CPS dominates every margin. At the scorer level a margin of 0.04 reaches 1.8% harm at regret 0.009, a point that CPS with the normal bound matches at $\delta \approx 0.2$ (1.8% harm, regret 0.010). The margin's harm is not controlled across testbeds, however (Section 5.3).

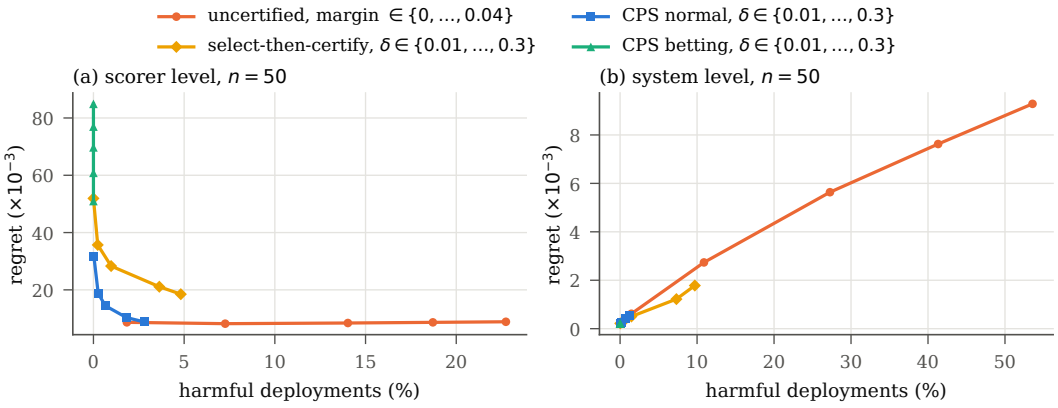

Figure 3: Harm–regret trade-off at $n = 50$ on Testbed A (MRR@10). Each curve varies δ (certified selectors) or the minimum-gain margin (uncertified selector).

Other utilities and tolerance. Table 4 reports, for each utility, harm, harm at tolerance 0.005, and regret at $n = 50$ and $n = 200$. The qualitative picture is the same for MAP@10 and Recall@10.

Table 4: Testbed A: harm (%) / harm by more than 0.005 (%) / regret ($\times 10^{-3}$), averaged over the 19 conditions at $n = 50$ and the 15 with $N \geq 400$ at $n = 200$.

Utility	Level	n	uncertified	t-test	CPS-norm	CPS-bet
MRR@10	scorer	50	22.9 / 13.5 / 9.1	5.7 / 3.5 / 9.9	0.7 / 0.6 / 13.3	0.0 / 0.0 / 70.8
		200	18.7 / 6.9 / 3.1	3.8 / 1.3 / 3.0	0.3 / 0.2 / 5.0	0.0 / 0.0 / 11.8
	system	50	53.8 / 42.1 / 9.6	9.6 / 7.7 / 1.9	0.2 / 0.2 / 0.3	0.0 / 0.0 / 0.2
		200	35.1 / 17.4 / 2.3	4.9 / 2.1 / 0.4	0.1 / 0.1 / 0.2	0.0 / 0.0 / 0.2
MAP@10	scorer	50	16.8 / 9.4 / 6.2	4.7 / 2.7 / 6.9	0.5 / 0.3 / 9.6	0.0 / 0.0 / 45.7
		200	17.1 / 4.9 / 2.4	3.9 / 1.1 / 2.4	0.3 / 0.2 / 4.4	0.0 / 0.0 / 12.9
	system	50	37.2 / 25.8 / 5.0	5.6 / 3.9 / 1.2	0.1 / 0.1 / 0.9	0.0 / 0.0 / 1.1
		200	24.7 / 8.3 / 1.2	3.0 / 0.9 / 0.2	0.1 / 0.0 / 0.2	0.0 / 0.0 / 0.2
Recall@10	scorer	50	22.8 / 18.0 / 8.9	5.5 / 4.5 / 9.3	0.6 / 0.5 / 13.8	0.0 / 0.0 / 23.7
		200	23.3 / 11.7 / 4.6	4.5 / 2.1 / 4.3	0.2 / 0.1 / 6.9	0.0 / 0.0 / 17.1
	system	50	28.8 / 20.7 / 4.6	4.6 / 3.4 / 1.1	0.2 / 0.2 / 0.6	0.0 / 0.0 / 0.6
		200	18.8 / 8.5 / 1.5	3.2 / 1.2 / 0.3	0.1 / 0.0 / 0.1	0.0 / 0.0 / 0.1

Testbed B per dataset. Table 5 reports every dataset with $N \geq 300$ at $n = 100$.

Table 5: Testbed B, the retriever's encoder adapting its own queries (MRR@10). "best ALL": best of 24 adaptive configurations; "#ALL>ret.": configurations that beat the retriever's order. Harm and deploy rates (%) over 400 calibration samples of $n = 100$ queries with $K = 48$ candidates.

					uncertified		t-test		CPS-norm	
Dataset	N	retriever	best ALL	#ALL>ret.	harm	deploy	harm	deploy	harm	deploy
ArguAna	1,401	0.297	0.336	17/24	0	100	0	100	0	82
NFCorpus	323	0.530	0.530	0/24	33	100	17	55	0	0
SciFact	300	0.685	0.681	0/24	25	100	10	52	0	0
SciDocs	1,000	0.334	0.327	0/24	67	100	36	57	0	0
HotpotQA	7,405	0.834	0.825	0/24	88	98	40	45	0	0
Quora	10,000	0.880	0.872	0/24	94	94	32	32	0	0
NQ	3,452	0.386	0.374	0/24	100	100	40	40	0	0
FiQA	648	0.465	0.451	0/24	90	98	28	32	0	0
FEVER	6,666	0.857	0.828	0/24	86	86	19	19	0	0

Budget prediction from a pilot. Estimating θ_c and σ_c from a pilot of 30, 50 or 100 labeled queries and plugging them into Equation (2) predicts the certification point within a factor of two for 42%, 49% and 59% of the 27 pairs from conditions with $N \geq 300$, and the pilot gap is non-positive in 16%, 12% and 7% of pilots. This is why Section 4 recommends sequential labeling with a time-uniform bound rather than a pilot-based plan.

E ANCHOR REGIMES

For each condition we split queries into entropy quintiles and, within each stratum, estimate ρ as the fraction of queries in which at least half of the anchors are relevant, together with γ and λ (Proposition 3). Of 95 strata with at least 30 queries and both anchor states present, 52 are in regime I, 28 in II, 14 in III and one in IV. With the top anchor alone defining Y, the counts are 49, 12, 11 and 23: the split between regimes II and IV depends on the definition of an informative anchor set, while regime I is stable. With Y as above, the majority regime (I: ALL; II: SKIP) gives the better of SKIP and ALL in all 17 conditions with $N \geq 300$ whose majority regime is I or II. Counting ROUTE as optimal when a population-fitted entropy gate beats the better simple action by more than 0.005, and predicting ROUTE for majority regime III, the majority regime is right in 16 of 19. It misses two Quora conditions whose routing gain comes from minority strata, and NQ with the retriever's encoder, whose three regime-III strata have $\rho < 0.06$, below $\rho^\star$. In the highest-entropy Quora slice, ρ is roughly constant (0.70–0.76), $\rho^\star$ falls from 0.99 to 0.33 across the regime-III quintiles, and μ changes sign between the third and the fourth quintile. This is the condition with the largest entropy headroom.

F Extended Related Work

Test-time adaptation and PRF. Test-time training and entropy minimization adapt a model on unlabeled test inputs (Sun et al., 2020; Wang et al., 2021a; Zhang et al., 2022). EATA and SAR update only on samples judged reliable (Niu et al., 2022; 2023), and TTA gains depend on hyperparameters that cannot be tuned without target labels (Zhao et al., 2023). In our episodic setting the offset is reset after each query, so a reliability filter acts as a per-query gate, whose added value Proposition 2 bounds. Dense PRF helps on some collections and hurts on others (Yu et al., 2021; Wang et al., 2021b; Li et al., 2023); query performance prediction tries to anticipate this without labels (Cronen-Townsend et al., 2002; Zhou & Croft, 2007; Shtok et al., 2012).

Treatment choice and deferral. A Route gate is a treatment rule. Proposition 2 is the value of the first-best rule on z (Manski, 2004), and empirical welfare maximization bounds the regret of estimated rules from observational data (Kitagawa & Tetenov, 2018; Athey & Wager, 2021); Lemma 1 is the simplest such bound, sufficient here because both outcomes are observed. Selective prediction and learning to defer (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Mozannar & Sontag, 2020) abstain or defer to an expert; here deferral routes to another ranking action whose outcome is observable offline.

Safe policy improvement. Deploying a policy only when a confidence bound certifies improvement over a baseline is the principle of high-confidence and safe policy improvement (Thomas et al., 2015; Laroche et al., 2019); learn-then-test (Angelopoulos et al., 2025) and off-policy evaluation (Dudík et al., 2011) are related. We use empirical-Bernstein and betting bounds and confidence sequences (Maurer & Pontil, 2009; Waudby-Smith & Ramdas, 2024; Howard et al., 2021); the label budget parallels best-arm identification (Kaufmann et al., 2016) and power analysis in IR evaluation, where multiple comparisons are a known pitfall (Webber et al., 2008; Sakai, 2016; Carterette, 2012). Active testing and prediction-powered inference could reduce the label cost further (Kossen et al., 2021; Angelopoulos et al., 2023). Our contribution is to cast TTA deployment as this problem, bound what gated adaptation can gain, and measure how often uncertified and uncorrected selection fail.