HARMCAP: STATE-AWARE CONSEQUENCE CONTROL FOR RETRIEVAL-CONDITIONED TABULAR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-conditioned tabular models use neighboring examples or in-context demonstrations to adapt predictions at inference time, but a strong retrieval mechanism does not by itself control how much one realized update may deteriorate an individual prediction. We introduce HarmCap, a post-retrieval controller that treats the retrieval-conditioned output as a proposal and selects the largest query-specific step along its realized logit-space direction whose worst label-wise cross-entropy increase stays within a prescribed budget. The admissible step depends on the full predictive state, making scalar confidence insufficient for exact multiclass control. Across public tabular benchmarks and multiple retrieval-conditioned interfaces, HarmCap satisfies the prescribed pointwise contract while retaining more of the proposed update than a state-agnostic exact-safe envelope. Protocol-aligned TabR experiments further recover the predictive behavior reported in the original work before applying HarmCap to matched same-checkpoint retrieval updates. The experiments also show that admissibility and predictive utility are distinct objectives, supporting consequence control as a separate inference-time component of retrieval-conditioned prediction.

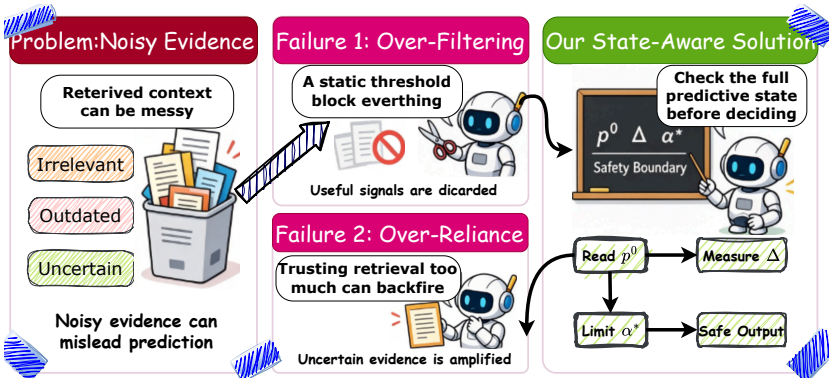

Figure 1: **Motivation.** The information retrieved in general may be useful or unreliable: Overly aggressive filtering may discard useful evidence, while adopting without restrictions may amplify harmful updates. However, HarmCap uses the current predictive state to control how many retrieval-induced updates are accepted.

1 INTRODUCTION

Modern tabular learning is increasingly moving beyond task-specific predictors toward retrieval-based and in-context inference. Earlier work established strong neural baselines while showing that deep models do not uniformly dominate tree-based methods across heterogeneous tabular problems (Gorishniy et al., 2021; Grinsztajn et al., 2022; McElfresh et al., 2023). More recent approaches revisit instance-level inductive biases (Ye et al., 2025), improve neural ensembling (Gorishniy et al., 2025), and develop tabular foundation models such as TabPFN (Hollmann et al., 2023; 2025), TabICL (Qu et al., 2025), and TabDPT (Ma et al., 2025). Large-scale benchmarks and surveys further reflect this shift toward transferable and inference-time adaptive prediction (Liu et al., 2025; Jiang et al., 2026).

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.

Retrieval makes this adaptation explicit. TabR augments learned representations with neighboring training instances and labels (Gorishniy et al., 2024), while retrieval and local fine-tuning can strengthen in-context tabular predictors (Thomas et al., 2024). TabICL and TabDPT likewise condition predictions on examples available at inference time (Qu et al., 2025; Ma et al., 2025). Yet retrieval quality and the consequence of a realized retrieval update are different questions. Indeed, TabR itself observes that retrieval is not uniformly beneficial across datasets (Gorishniy et al., 2024). Even when retrieval improves average predictive performance, an individual query can still undergo a large or harmful prediction change. This motivates a complementary question: *how much influence should a realized retrieval-induced update be allowed to exert on one prediction?*

Existing reliability mechanisms address related but different objects. Retrieval-side approaches improve which evidence is used through self-reflection, robustness to irrelevant context, sufficiency assessment, pruning, or retrieval-utility estimation (Asai et al., 2024; Yoran et al., 2024; Dai et al., 2025; Joren et al., 2025; Chirkova et al., 2025; Huang et al., 2025; Shen et al., 2025). Output-side methods instead constrain quantities such as policy divergence, expected risk, prediction-set coverage, or generation risk (Schulman et al., 2015; Angelopoulos et al., 2024; H. Zargarbashi et al., 2024; Kang et al., 2024; Xu et al., 2025). These methods provide useful safeguards, but they do not directly impose a query-specific, label-free bound on the worst classification-loss increase caused by accepting a fixed tabular retrieval update.

We study this problem as *post-retrieval consequence control*. Given a base prediction and a retrieval-conditioned proposal, the goal is to retain as much of the proposed update as possible while limiting its worst possible consequence before the query label is observed. We introduce **HarmCap**, a lightweight controller that preserves the realized logit-space update direction and selects the largest query-specific admissible step under a prescribed consequence budget. Its key property is state dependence: the consequence of an update depends jointly on the proposed direction and the current predictive distribution. As a result, updates with similar magnitude—or predictions with similar scalar confidence—can admit substantially different safe steps.

Our contributions are threefold. First, we formulate *post-retrieval consequence control* as a distinct inference-time problem that separates retrieval proposal generation from pointwise admissibility control and optional usefulness selection. Second, we derive HarmCap as the exact maximal controller on the realized logit-space update ray, characterize its dependence on the full predictive state, and validate the resulting solution with high-precision numerical and maximality audits. Third, we evaluate the method on established tabular benchmarks and retrieval interfaces, including a protocol-aligned reproduction of TabR on its original classification benchmarks, TALENT, OpenML-CC18, TabZilla, TabICL, and same-checkpoint TabR updates. Comparisons with state-agnostic and alternative exact-safe controllers further reveal that admissibility, proposal preservation, and predictive utility are distinct objectives.

Detailed comparisons with related formulations are provided in Appendix A.

2 HarmCap

HarmCap operates as a post-retrieval control interface between an existing predictor and its final output. Given a base prediction and a retrieval-conditioned prediction defined on the same ordered class support, it determines how much of the realized retrieval-induced update can be accepted under a prescribed per-query consequence budget. The underlying retriever and predictor remain unchanged; HarmCap acts only on their outputs at inference time.

2.1 From a retrieval proposal to a pointwise consequence constraint

For query q, let $z_q^0, z_q^r \in \mathbb{R}^K$ denote the base and retrieval-conditioned logits on the same ordered class support. Their difference defines the realized retrieval proposal,

$$\Delta_q = z_q^r - z_q^0, \qquad z_q(\alpha) = z_q^0 + \alpha\Delta_q, \qquad \alpha \in [0,1]. \tag{1}$$

Here $\alpha = 0$ recovers the base prediction and $\alpha = 1$ recovers the full retrieval-conditioned proposal. For cross-entropy $\ell_y(z) = -z_y + \mathrm{LSE}(z)$, we define the consequence of taking step α as the largest loss increase over all possible labels,

$$H_q(\alpha) = \max_y \left\{ \ell_y(z_q(\alpha)) - \ell_y\left(z_q^0\right) \right\}. \tag{2}$$

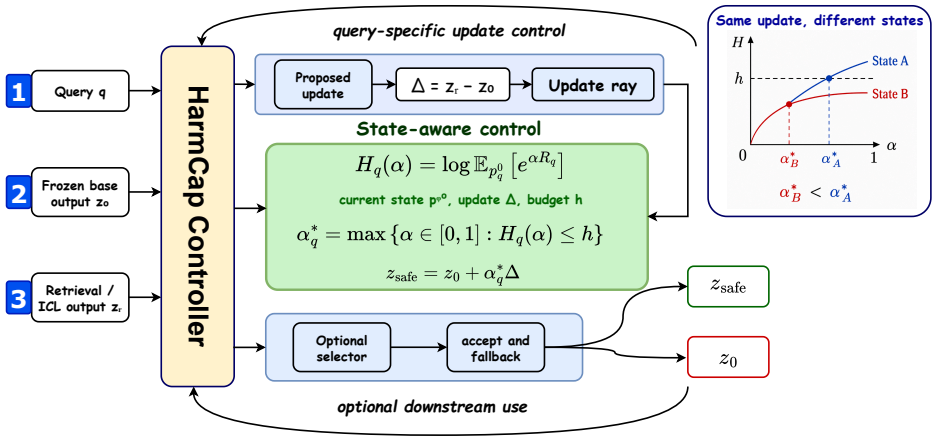

Figure 2: **HarmCap as post-retrieval consequence control.** The base and retrieval-conditioned outputs define a realized prediction update. HarmCap uses the current predictive state and a prescribed consequence budget to determine the maximal admissible step along that update direction. Proposal generation and optional downstream usefulness selection remain separate components.

An update is admissible when $H_q(\alpha) \le h$, where $h \ge 0$ is specified before test evaluation. Because the maximum is taken over all labels, admissibility does not require access to the unknown test label.

Let $p_q^0 = \mathrm{softmax}(z_q^0)$ and $p_q^\alpha = \mathrm{softmax}(z_q(\alpha))$. Under softmax cross-entropy,

$$H_q(\alpha) = D_\infty\left(p_q^0 \| p_q^\alpha\right) = \max_y \log \frac{p_q^0(y)}{p_q^\alpha(y)}. \tag{3}$$

Therefore the contract $H_q(\alpha) \le h$ is equivalent to

$$p_q^\alpha(y) \ge e^{-h} p_q^0(y), \qquad \forall y. \tag{4}$$

Thus h has a direct probability-ratio interpretation: no class probability is allowed to fall below an e^{-h} fraction of its base value. For example, $h = 0.1$ implies $e^{-0.1} \approx 0.905$, so every class probability remains at least approximately 90.5% of its base value. This interpretation is independent of the number of classes, although the admissible step itself depends on the full predictive state. Proofs and equivalent forms are given in Appendix B.

2.2 Maximal admissible same-ray control

HarmCap chooses the largest coefficient on the realized retrieval ray that satisfies the consequence contract,

$$\alpha_q^\star = \max\{\alpha \in [0,1] : H_q(\alpha) \le h\}, \qquad z_q^{\text{safe}} = z_q^0 + \alpha_q^\star \Delta_q. \tag{5}$$

To see why this one-dimensional problem is well posed, define

$$R_{q,y} = \Delta_{q,y} - \min_j \Delta_{q,j} \ge 0.$$

Then

$$H_q(\alpha) = \log \sum_y p_q^0(y) e^{\alpha R_{q,y}} = \log \mathbb{E}_{Y \sim p_q^0}\left[e^{\alpha R_q(Y)}\right]. \tag{6}$$

Hence H_q is a log-moment-generating function of a nonnegative random variable under the current predictive state. It follows that $H_q(0) = 0$ and that $H_q(\alpha)$ is continuous, nondecreasing, and convex on $[0,1]$. Consequently,

$$\{\alpha \in [0,1] : H_q(\alpha) \le h\} = [0, \alpha_q^\star] \tag{7}$$

for some unique maximal boundary point whenever the full proposal is inadmissible. This prefix-interval structure makes $\alpha_q^\star$ recoverable with a one-dimensional root search.

Using the identity above, HarmCap evaluates

$$H_q(\alpha) = \mathrm{LSE}(z_q^0 + \alpha \Delta_q) - \mathrm{LSE}(z_q^0) - \alpha \min_j \Delta_{q,j}, \tag{8}$$

which can be computed stably without explicitly forming loss values for every hypothetical label. If

Algorithm 1 HarmCap numerical core

Require: Matched logits $z^0, z^r \in \mathbb{R}^K$, budget h, iterations B
$\Delta \leftarrow z^r - z^0$
Define $H(a) \leftarrow \mathrm{LSE}(z^0 + a\Delta) - \mathrm{LSE}(z^0) - a \min_j \Delta_j$
if $H(1) \le h$ **then**
  return $(z^r, 1)$
end if
$(a_{\mathrm{lo}}, a_{\mathrm{hi}}) \leftarrow (0, 1)$
for $b = 1, \ldots, B$ **do**
  $a \leftarrow (a_{\mathrm{lo}} + a_{\mathrm{hi}})/2$
  if $H(a) \le h$ **then**
    $a_{\mathrm{lo}} \leftarrow a$
  else
    $a_{\mathrm{hi}} \leftarrow a$
  end if
end for
return $(z^0 + a_{\mathrm{lo}}\Delta, a_{\mathrm{lo}})$

$H_q(1) \le h$, the full retrieval proposal is accepted. Otherwise, HarmCap bisects the interval $[0, 1]$ and returns the largest feasible coefficient.

Each evaluation of H_q is $O(K)$, giving $O(BK)$ controller complexity after the base and retrieval-conditioned predictions are available. The released implementation uses float64 arithmetic and $B = 64$ bisection iterations. The mathematical feasibility test is $H_q(\alpha) \le h$; the separate evaluation audit records a numerical violation when $H > h + 10^{-8}$. Section 3.4 compares the implementation against high-precision root solutions and explicit maximality checks. All formal evaluations use float64 because pure float32 is not numerically reliable for decisions extremely close to the admissibility boundary.

2.3 Control geometry

The consequence contract specifies which predictions are admissible, but it does not by itself specify how to move within that admissible set. HarmCap adopts a direction-preserving geometry: it keeps the realized retrieval-induced direction fixed in logit space and modifies only its magnitude. Accordingly, Eq. (5) solves a maximal-step problem on the path

$$\mathcal{P}_q^{\mathrm{logit}} = \left\{ z_q^0 + \alpha(z_q^r - z_q^0) : \alpha \in [0, 1] \right\}. \tag{9}$$

This design preserves the update proposed by the underlying retrieval-conditioned predictor rather than replacing it with a newly optimized prediction. It also separates two modeling choices: the *consequence contract*, which defines admissibility, and the *control path*, which defines which admissible predictions are reachable by the controller. Alternative paths between the same base and retrieval-conditioned outputs may therefore produce different proposal-preservation and predictive-utility trade-offs. We treat such alternatives as controller baselines in the experiments; HarmCap specifically targets exact maximal control on the realized logit-space proposal.

2.4 Why the current predictive state matters

Equation (6) shows that the same classwise displacement can have different consequences under different base predictive states: the probabilities $p_q^0(y)$ weight the exponentiated relative displacement $R_{q,y}$. Thus two queries with the same logit update Δ_q need not admit the same coefficient $\alpha_q^\star$.

A state-agnostic exact-safe envelope follows from

$$H_q(\alpha) \le \alpha \ \mathrm{range}(\Delta_q), \qquad \alpha_q^{\mathrm{clip}} = \min\left\{1, \frac{h}{\mathrm{range}(\Delta_q)}\right\} \le \alpha_q^\star. \tag{10}$$

Range Clip is therefore guaranteed safe under the same consequence contract, but it can be conservative because it ignores how the current predictive state distributes probability mass across classes.

Compressed summaries of the predictive state can also be insufficient. In binary classification, the identity and probability of the top class determine the full predictive distribution. For $K > 2$, however, the remaining probability mass can be distributed among non-top classes in different ways while preserving the same top-class confidence. Because Eq. (6) depends on the complete set of probabilities $\{p_q^0(y)\}_{y=1}^K$, such matched-confidence states can have different admissible radii. Appendix B gives an explicit construction, and Section 3.3 evaluates this effect with confidence-only and strict partial-state diagnostics on TALENT-MC80.

3 EXPERIMENTS

We evaluate HarmCap across public tabular benchmarks and multiple retrieval-conditioned prediction interfaces, with experiments designed to test three questions: whether the controller satisfies the prescribed pointwise consequence budget, how much of the retrieval-induced proposal it preserves relative to alternative safe controllers, and how these guarantees interact with predictive utility and downstream usefulness.

3.1 DATASETS, INTERFACES, METRICS, AND PUBLISHED ALIGNMENT

TALENT provides unified tabular datasets and implementations (Liu et al., 2025). TALENT-MC80 contains all 80 official multiclass tasks ($K = 3$–100; 279,852 test queries) and is our primary full-state mechanism/controller board. External replication uses all 72 OpenML-CC18 tasks (Bischl et al., 2021) and 35/36 TabZilla hard-suite tasks (McElfresh et al., 2023); the remaining TabZilla task, `albert` (189356), fails with a pre-result OOM. The three public suites use the same logistic–cosine-kNN proposal recipe ($k = 20$, retrieval coefficient $\eta = 0.25$, Laplace smoothing $\varepsilon = 1$). Cross-substrate evidence additionally uses a pre-stratified 16-task TabICL kNN-context interface and a 16-task CC18 TabR same-checkpoint ablation in which the retrieval contribution `context_x` is set to zero for z^0 while the shared downstream blocks and prediction head are unchanged. The TabICL context path is author-constructed, and the TabR pair is a LEVEL-B same-checkpoint ablation rather than an official retrieval-off API. Finally, CC18 and TabZilla replay six pre-specified budgets $h \in \{.01, .025, .05, .1, .2, .5\}$ without retraining the predictors.

To anchor the controller evaluation to published tabular retrieval rather than to an author-defined substrate alone, we first evaluate TabR (Gorishniy et al., 2024) on four completed classification datasets from its original public benchmark (CH, AD, OT, and HI). We use the published fixed splits and Accuracy metric, the official published/best configurations, and 15 seeds without hyperparameter retuning. This is a protocol-aligned LEVEL-P2 reproduction rather than an exact LEVEL-P1 rerun of the full tuning procedure. The resulting mean accuracies differ from the paper by at most 0.0011 in absolute terms. We then freeze same-checkpoint retrieval-on/retrieval-disabled prediction pairs on these datasets and apply Range Clip and HarmCap to the realized updates. Covertype (CO), the fifth classification dataset in the original TabR benchmark, was still running at the artifact freeze and is not used in any claim. Table 3 separates the published predictor comparison from the controller overlay.

To test whether the empirical conclusions depend on a weak comparator, we additionally evaluate three strong label-free safe controllers on the same frozen (z^0, z^r) pairs at $h = 0.1$. *Moment Envelope* stays on the realized logit ray but replaces the exact log-MGF with a moment-based sufficient upper envelope; *Probability-Mix-Max* instead follows the probability-space segment $(1-\beta)p^0 + \beta p^r$ and takes the largest closed-form β satisfying the same classwise probability lower bound. The latter is maximal in a different geometry, so its norm-ratio Retention is not directly comparable to same-ray Retention. All methods use the same task sets, budget, and 10^{-8} evaluation tolerance, with no test-label tuning. We report Raw retrieval as an unconstrained predictive reference rather than as a safe baseline.

The primary controller metrics are Retention, beneficial-update retention (BUR), exact-H violations, and the observed maximum H. Retention is the released norm ratio

$$\mathrm{Ret}_q = \frac{\|z_q^{\mathrm{ctrl}} - z_q^0\|_2}{\max(\|z_q^r - z_q^0\|_2, 10^{-12})}, \tag{11}$$

which equals the applied coefficient for a non-negligible same-ray update. BUR is label-dependent and evaluation-only (Appendix E.1). Standard predictive effects are measured with NLL, Brier score, and balanced accuracy (Gneiting & Raftery, 2007; Brodersen et al., 2010). We report task-macro

means and paired task-bootstrap 95% intervals (10,000 replicates; seed 20260904). Reproducibility closure regenerates the CC18, TabZilla, and TabICL16 headline values from canonical matched-pair archives and metric code. Algorithmic validation compares the released 64-step float64 bisection with high-precision reference roots on 36,407 sampled real queries across five suites/interfaces; the maximum observed coefficient discrepancy is 9.4×10^{-15}. A separate synthetic audit retains 58 approximately-10^{-8} boundary cases, so we describe the mathematical same-ray solution as exact while treating the floating-point implementation with an explicit evaluation tolerance rather than as a strict numerical certificate. Full details are in Appendix D.

3.2 Main results: external replication and controller families

The HarmCap–Range Clip ordering transfers across datasets and proposal interfaces, but its absolute magnitude is substrate-dependent. The largest Retention gains occur on the logistic–kNN boards, whereas the native TabR closure is smaller: Retention rises from 15.3% to 21.1%, ΔRet$= +5.79$ pp [3.28, 8.82], with 16/0/0 task-level wins and zero HarmCap violations. This matters because the TabR pair changes only the retrieval contribution within the same checkpoint rather than reusing the logistic–kNN proposal family. Standard predictive effects are heterogeneous: on CC18 and TabZilla, mean NLL/Brier changes relative to Range Clip are small on average and many individual tasks move in the opposite direction. Consequently, cross-substrate transfer is claimed for consequence control and proposal preservation under a fixed protocol, not for a universal Retention level or predictive improvement.

The fair board changes the interpretation of the empirical claim. HarmCap remains substantially less conservative than Range Clip and improves native-TabR NLL by -0.061 with a paired 95% interval $[-0.110, -0.024]$, but it is not the uniquely best predictive safe controller. Moment Envelope nearly matches HarmCap on the same ray (NLL 0.521 versus 0.518 and Retention 0.210 versus 0.211), while Probability-Mix-Max achieves lower NLL (0.503) by moving along a different probability-space path. Raw retrieval is predictively strongest on this board but violates the $h = 0.1$ contract on about 91% of queries. Across CC18, TabZilla, TabICL16, and the four TabR-original datasets, no single safe geometry uniformly minimizes NLL. We therefore use HarmCap's exact same-ray maximality—not universal predictive dominance—as the method claim; the complete cross-suite fair board is reported in Appendix I.

We further anchor HarmCap to the original TabR benchmark. Under a 15-seed protocol-aligned evaluation, our TabR implementation closely recovers the published accuracy on all four completed classification datasets. Using the same fixed checkpoints, we then evaluate HarmCap on the resulting retrieval-induced updates. HarmCap consistently retains more of each update than Range Clip while maintaining zero exact-H violations across CH, AD, OT, and HI. This provides a direct validation of consequence control on top of an established retrieval-conditioned tabular predictor rather than on a custom retrieval substrate alone.

3.3 State dependence and ablations

The partial-state experiments support the state-dependence predicted by Eq. (6). Confidence-only reconstruction violates the exact consequence budget on 3.94% of MC80 queries, and none of the evaluated strict top-k representations eliminates violations on its eligible tasks. The top-k results are diagnostic rather than a monotone ablation because the eligible task set changes with k. Complementary residual-mass interventions, which preserve the top-class confidence while redistributing probability among non-top classes, shift the admissible radius by 0.28–0.66 across the predeclared class-count groups (Appendix C). Together, these results show that exact multiclass control depends on information beyond scalar confidence. They do not imply a monotone dependence on the number of classes. The corresponding MC80 utility interval crosses zero, further emphasizing that admissibility and predictive benefit are separate properties.

3.4 Budget sensitivity and implementation

Numerical accuracy. We validate the numerical solver against high-precision root finding and explicit maximality checks on real-query pairs. The 64-step float64 implementation recovers the maximal admissible same-ray coefficient to approximately machine precision on the audited queries.

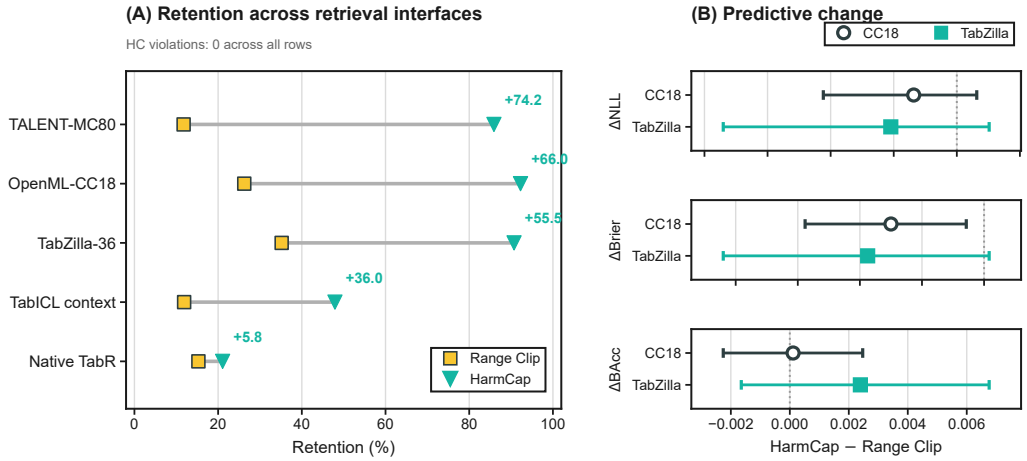

Figure 3: External replication. (A) HarmCap–Range Clip paired Retention gains across three public logistic–kNN suites and a second TabICL context substrate; horizontal segments are task-bootstrap 95% CIs. HarmCap records zero exact-H violations on every completed task. The later native-TabR same-checkpoint closure is reported in Table 1. (B) Standard predictive-score changes on CC18 and TabZilla. The intervals support a large retention advantage without a predictive-gain claim.

Table 1: Headline safety–efficiency results at $h = 0.1$. Retention/BUR are percentages and paired gains are percentage points. TabICL is an author-constructed context interface; TabR is a LEVEL-B same-checkpoint retrieval ablation (`context_x` $= 0$ for the base path), not an official retrieval-off API. Yellow highlights HarmCap rows, not predictive dominance.

Method	Ret. ↑	BUR ↑	Viol. # ↓	ΔRet. [95% CI]
TALENT-MC80 / logistic–kNN — 80/80 tasks; 279,852 queries				
Range Clip	11.75	9.84	0	—
HarmCap	**85.91**	**82.37**	**0**	**+74.15 [69.91, 78.28]**
OpenML-CC18 / logistic–kNN — 72/72 tasks; 87,494 queries				
Range Clip	26.24	21.46	0	—
HarmCap	**92.26**	**85.34**	**0**	**+66.02 [60.14, 71.71]**
TabZilla-36 / logistic–kNN — 35/36 tasks; 201,934 queries				
Range Clip	35.18	27.38	0	—
HarmCap	**90.71**	**84.23**	**0**	**+55.53 [47.70, 63.29]**
External TabICL kNN context — 16/16 pre-stratified tasks				
Range Clip	11.88	11.73	0	—
HarmCap	**47.91**	**27.34**	**0**	**+36.04 [23.86, 49.25]**
Native TabR / CC18 subset — 16/16 tasks; LEVEL-B same-checkpoint ablation				
Range Clip	15.30	9.10	0	—
HarmCap	**21.10**	**26.00**	**0**	**+5.79 [3.28, 8.82]**

Table 2: Fair controller comparison on the 16-task native TabR same-checkpoint board at $h = 0.1$. Raw is an unconstrained reference. Probability-Mix-Max is maximal on a probability-space segment, so the logit norm-ratio Retention used for same-ray controllers is not comparable and is shown as "–". Runtime is controller-only CPU latency from the common $n = 1024, K = 5$ synthetic audit and excludes TabR inference and retrieval construction.

Controller	Control geometry	Ret.	NLL ↓	Brier ↓	Bal. Acc. ↑	Viol. ↓	μs/query ↓
Base	no update	0.000	0.596	0.323	0.712	0.000	0.002
Raw	unconstrained proposal	1.000	**0.315**	**0.178**	**0.823**	0.911	0.003
Range Clip	state-agnostic logit envelope	0.153	0.579	0.314	0.718	**0.000**	0.270
Moment Envelope	moment bound on realized ray	0.210	0.521	0.296	**0.725**	**0.000**	**0.147**
Prob.-Mix-Max	maximal probability segment	–	**0.503**	**0.291**	0.725	**0.000**	0.196
HarmCap	**exact maximal realized logit ray**	**0.211**	0.518	0.296	**0.725**	**0.000**	28.660

Because decisions extremely close to the boundary remain sensitive to finite precision, we do not interpret the implementation as a strict floating-point certificate. All formal experiments therefore use float64 together with the declared evaluation tolerance. Full precision and stress-test results are reported in Appendix D.

Computational cost. HarmCap evaluates the $O(K)$ consequence function repeatedly, giving $O(BK)$ controller complexity for B bisection steps. Its iterative solver is therefore slower than the

Table 3: **Published TabR alignment and matched consequence control.** (A) Protocol-aligned Accuracy comparison with TabR ICLR'24. (B) Consequence control on matched same-checkpoint retrieval updates.

(A) TabR alignment				(B) Matched consequence control			
Dataset	Paper	Ours	Gap	Dataset	Clip Ret.	HC Ret.	HC Viol.
CH	.862	.8626±.0028	.0006	CH	8.41	**15.43**	**0.00**
AD	.870	.8700±.0015	.000008	AD	29.84	**59.01**	**0.00**
OT	.825	.8247±.0020	.000259	OT	1.90	**4.61**	**0.00**
HI	.729	.7301±.0017	.0011	HI	16.78	**31.99**	**0.00**

Table 4: **State dependence on TALENT-MC80.** Panel A compares state-agnostic, compressed-state, and full-state controllers under the same headline budget $h = 0.1$. Panel B progressively restricts the predictive state to its top-k components. The strict top-k diagnostic uses the complete state only to evaluate the exact consequence; because the eligibility requirement depends on k, the task sets are not identical across rows.

A. How much predictive state is needed for exact control?				
Controller	State information	Ret. (%)	$H_{\max}$	Viol. (%)
Range Clip	State-agnostic envelope	11.75	0.0932	**0.00**
Confidence-only	Scalar confidence	44.90	0.1646	3.94
HarmCap	**Full predictive state**	**85.91**	**0.1000**	**0.00**
B. Strict partial-state reconstruction diagnostic				
State available	Interpretation	Tasks	Ret. (%)	Viol. (%)
Top-1	Largest class only	80	44.9	3.98
Top-2	Two largest classes	80	60.4	1.43
Top-3	Three largest classes	55	57.8	2.22
Top-5	Five largest classes	39	61.2	0.92
Top-10	Ten largest classes	12	40.5	0.46
Full	**Complete class state**	80	**85.9**	**0.00**

closed-form safe controllers considered in our comparison, although the measured cost remains small relative to typical retrieval and predictor inference. Detailed controller-only runtimes are reported in Appendix J.

3.5 Downstream usefulness under shared gates

Admissibility alone does not determine whether a retrieval-induced update is predictively useful. We therefore apply the same predefined, label-free gates to every controller on each frozen prediction pair, with rejected queries falling back to the base prediction. Because the accept mask is shared across controllers and is fixed without test labels, the comparison isolates the controlled output rather than a controller-specific selection rule.

The resulting picture is consistent across the evaluated interfaces. HarmCap retains an NLL advantage over Range Clip after selection on Native TabR and the published-aligned TabR boards, whereas the corresponding difference on TabICL is negligible. All gated comparisons continue to satisfy the prescribed exact-H contract. We therefore treat usefulness selection as complementary to consequence control: HarmCap determines how far an admissible proposal may move, while a separate selector determines whether that update should be used. Gate definitions and complete paired results are provided in Appendix I.

4 Discussion and limitations

HarmCap separates two questions that are often conflated in retrieval-conditioned prediction: whether external information should be used, and how strongly a realized retrieval-induced change should be allowed to affect an individual prediction. The former concerns retrieval quality or usefulness; the latter is the consequence-control problem studied here. HarmCap addresses this second question by conditioning the admissible step on the current predictive state and the realized update direction. This perspective explains why state-agnostic envelopes can be unnecessarily conservative and why compressed summaries such as scalar confidence are insufficient for exact multiclass control. More broadly, our results suggest that retrieval-conditioned prediction benefits from treating proposal generation and consequence control as distinct components rather than asking a single retrieval score to serve both roles.

Table 5: Algorithmic validation summary. Numerical agreement establishes the stated same-ray solution but is not presented as an all-regime floating-point certificate.

Check	Evaluation	Result	Interpretation
FP64 solver vs. high-precision root	36,407 real queries	max $\|\Delta\alpha\| = 9.4 \times 10^{-15}$	reference agreement
Maximality on realized ray	contracted real queries	ref. agreement 100%; $+10^{-8}$ crosses boundary	maximal same-ray step
Iteration study	pooled reference set	B=16: p99 1.5×10^{-5}; B=24: 5.8×10^{-8}	B=64 is conservative
Finite-precision stress	FP32 + synthetic audit	false-safe boundary cases; 58 tiny excesses retained	use declared tolerance
Budget semantics	$h = .01$–$.5$	classwise floor e^{-h} of base probability	application-facing interpretation

Table 6: **Offline utility headroom under the same consequence budget.** "Ray–HC" is the label-aware gain available on the realized ray; "Full–ray" is the additional gain from changing the update direction. Both are offline, non-deployable oracle quantities.

Suite / interface	Published source	Ray–HC	Full–ray	Total
TALENT-MC80	TALENT, JMLR'25 (Liu et al., 2025)	.0462	**.1484**	.1945
OpenML-CC18	OpenML, NeurIPS D&B'21 (Bischl et al., 2021)	.0402	**.0611**	.1013
TabZilla-36	TabZilla, NeurIPS'23 (McElfresh et al., 2023)	.0364	**.0684**	.1047
TabICL16	TabICL, ICML'25 (Qu et al., 2025)	.0108	**.0466**	.0574
Native TabR	TabR, ICLR'24 (Gorishniy et al., 2024)	.0090	**.0573**	.0663

The broader controller comparison also reveals an important geometric aspect of the problem. HarmCap is exact for the particular path it controls: it returns the maximal admissible point along the realized logit-space update direction. This property should not be interpreted as global optimality over all admissible predictions. Different paths between the base and retrieval-conditioned outputs can induce different trade-offs between proposal preservation and predictive utility, and an admissible update need not be beneficial. Consequently, the choice of control geometry is itself a modeling decision. HarmCap deliberately preserves the predictor's proposed direction, which makes it lightweight and leaves the underlying retrieval system unchanged, but this same restriction limits its ability to correct a poor proposal. A separate usefulness mechanism can therefore be composed with consequence control when the goal is to decide whether an admissible update should ultimately be applied.

Several limitations follow from this formulation. First, HarmCap controls the magnitude of a fixed retrieval proposal rather than optimizing the proposal direction itself. Our oracle analysis indicates that relaxing this restriction can leave substantial additional utility headroom, motivating future controllers that jointly reason about direction and step size under the same consequence budget. Second, the present theory relies on finite-class softmax prediction with cross-entropy; extending the framework to regression, structured prediction, margin-based classifiers, or generative models will require consequence measures with different geometry. Third, the consequence budget h is application-dependent. Although it admits a direct multiplicative probability interpretation, selecting an appropriate operating point ultimately depends on the tolerance of the downstream application. Finally, the current implementation uses high-precision arithmetic near the admissibility boundary and incurs more computation than closed-form conservative alternatives. Faster solvers and adaptive-precision implementations are therefore natural engineering extensions.

Taken together, these results position HarmCap as an inference-time control interface for retrieval-conditioned prediction. HarmCap makes the consequence of a realized retrieval-induced update explicit and controllable, while proposal generation, utility assessment, and the choice of control geometry remain separate design decisions.

5 Conclusion

Overall, HarmCap provides a principled way to control the effect of retrieval at inference time. It complements retrieval and prediction by determining how much of a realized retrieval-induced update can be accepted under an explicit per-query consequence budget, while leaving proposal generation and usefulness assessment to separate components.

AI Assistance Statement

Generative AI tools were used to assist with literature discovery, code development and debugging, manuscript editing, figure preparation, and discussion of mathematical arguments. All AI-assisted

content was reviewed and verified by the authors against the relevant sources and experimental artifacts. The authors take full responsibility for the final manuscript.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. In *International Conference on Learning Representations*, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations*, 2024.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c7e1249ffc03eb9ded908c236bd1996d-Abstract-round2.html.

Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010. doi: 10.1109/ICPR.2010.764.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785–794*, 2016.

Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stephane Clinchant. Provence: Efficient and Robust Context Pruning for Retrieval-Augmented Generation. In *International Conference on Learning Representations*, 2025.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. SePer: Measure Retrieval Utility Through the Lens of Semantic Perplexity Reduction. In *International Conference on Learning Representations*, 2025.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems, 34*, 2021.

Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. TabR: Tabular Deep Learning Meets Nearest Neighbors. In *International Conference on Learning Representations*, 2024.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. In *International Conference on Learning Representations*, 2025.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Soroush H. Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17123–17147, 2024.

Noah Hollmann, Samuel Mueller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. To Trust or Not to Trust? Enhancing Large Language Models' Situated Faithfulness to External Contexts. In *International Conference on Learning Representations*, 2025.

Jun-Peng Jiang, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. Representation learning for tabular data: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6488–6508, 2026. doi: 10.1109/TPAMI.2026.3657217.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. In *International Conference on Learning Representations*, 2025.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models. In *Proceedings of the 41st International Conference on Machine Learning: 22963–23000*, 2024.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems, 30*, 2017.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, Huai-Hong Yin, Tao Zhou, Jun-Peng Jiang, and Han-Jia Ye. TALENT: A tabular analytics and learning toolbox. *Journal of Machine Learning Research*, 26 (226):1–16, 2025.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maks Volkovs. TabDPT: Scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, volume 36, pp. 76336–76369, 2023.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems, 31*, 2018.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847, 2025.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning: 1889–1897*, 2015.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. ReliabilityRAG: Effective and Provably Robust Defense for RAG-based Web-Search. In *Advances in Neural Information Processing Systems, 38*, 2025.

Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony Caterini. Retrieval & Fine-Tuning for In-Context Tabular Models. In *Advances in Neural Information Processing Systems, 37*, 2024.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback–leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. A Theory for Token-Level Harmonization in Retrieval-Augmented Generation. In *International Conference on Learning Representations*, 2025.

Han-Jia Ye, Huai-Hong Yin, De-Chuan Zhan, and Wei-Lun Chao. Revisiting Nearest Neighbor for Tabular Data: A Deep Tabular Baseline Two Decades Later. In *International Conference on Learning Representations*, 2025.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *International Conference on Learning Representations*, 2024.

APPENDIX CONTENTS

A Detailed related work p. 13
B Proofs and identities p. 13
C MC80 state–update mechanism details p. 14
D Algorithmic and numerical validation p. 15
E Reproducibility, protocols, and metric definitions p. 16
F TabR-36 frozen selection summary p. 17
G Predictive breadth and published-model sanity checks p. 17
H External tabular validation p. 17
I Additional retrieval-interface and risk-object evidence p. 19
J Budget sensitivity and computational scope p. 20
K Limitations p. 20

A DETAILED RELATED WORK

Prediction interfaces versus output control. TabR incorporates neighbors into a learned tabular predictor (Gorishniy et al., 2024); ModernNCA learns representations for effective nonparametric prediction (Ye et al., 2025). TabM instead strengthens parametric tabular models through parameter-efficient ensembling (Gorishniy et al., 2025). These differences matter when constructing a meaningful base/context pair: a strong standalone predictor is not automatically a retrieval-conditioned control interface. We therefore report published predictive scores separately from HarmCap comparisons.

TabPFN uses prior-data fitted in-context prediction (Hollmann et al., 2023; 2025), and TabICL separates within-row representation from interaction between examples to scale this paradigm (Qu et al., 2025). TabDPT studies pretraining on real tabular data (Ma et al., 2025). Thomas et al. (2024) introduce retrieval and local fine-tuning for in-context tabular models. HarmCap assumes their output is already available and does not replace their training objective, context selection, or adaptation procedure. In particular, LoCalPFN comparisons using two model states are labeled cross-model rather than silently presented as retrieval-on/off ablations.

Reliability of retrieved evidence. Self-RAG learns when to retrieve and how to critique generated content (Asai et al., 2024). Yoran et al. (2024) use filtering and training with mixed relevant/irrelevant evidence to reduce context-induced failures. Sufficient Context distinguishes a failure of evidence sufficiency from a model's failure to use adequate evidence (Joren et al., 2025). Provence performs sentence-level pruning and context reranking (Chirkova et al., 2025), while Huang et al. (2025) study situated faithfulness when external context conflicts with internal knowledge. These methods target evidence use and answer quality; our tabular experiments do not evaluate their language-model implementations. The methodological distinction is that HarmCap tests the realized probability update after context processing, with a specified relative-loss budget.

Benefit, detriment, and different risk objects. Xu et al. (2025) analyze distribution fusion in RAG and propose Tok-RAG for token-level harmonization. That benefit–detriment analysis is closely related motivation, but is not the finite-support, worst-label maximal-ray-step problem evaluated here. We do not claim that its formulas fail in multiclass settings, nor that the standard Rényi divergence is new.

Conformal Risk Control calibrates expected losses under its assumptions (Angelopoulos et al., 2024); robust conformal sets use bounds on conformity-score perturbations to protect coverage (H. Zargarbashi et al., 2024). Trust-region policy optimization regulates policy movement through a KL-based constraint (Schulman et al., 2015). We use KL and finite-order Rényi *ray controllers* as reference adaptations, not as reproductions of those full algorithms. Their scalar budgets are selected on development data to match the empirical exact-H target. The mathematical divergences follow Kullback & Leibler (1951) and van Erven & Harremoës (2014). Exact-H violations in our experiments do not imply failure of a comparator's own risk guarantee.

Benchmark scope. TALENT supplies a unified toolbox and dataset collection (Liu et al., 2025). OpenML represents datasets, targets, and evaluation tasks (Vanschoren et al., 2013); the CC18 suite standardizes a curated classification collection (Bischl et al., 2021). TabZilla concentrates on tasks found difficult in a broad model comparison (McElfresh et al., 2023). Reusing these public suites improves transparency of the evaluation setting, but distinct suite names alone do not establish non-overlapping datasets. We report the suites separately and do not sum their counts into a claim about unique external datasets.

B PROOFS

This appendix collects the identities used in Section 2. None of them introduces a new divergence or a new MGF theory; they specialize standard facts to the HarmCap ray.

B.1 $H = D_\infty$

Let $\ell_y(z) = -z_y + \mathrm{LSE}(z)$ with $\mathrm{LSE}(z) = \log \sum_j e^{z_j}$ and $p = \mathrm{softmax}(z)$. Then

$$\ell_y(z+\Delta) - \ell_y(z) = \log \frac{p(y)}{\mathrm{softmax}(z+\Delta)(y)},$$

so $H_q(\alpha) = \max_y \log(p_q^0(y)/p_q^\alpha(y)) = D_\infty(p_q^0 \| p_q^\alpha)$ (van Erven & Harremoës, 2014).

B.2 LOG-MGF CHARACTERIZATION

Let $m_q = \min_j \Delta_{q,j}$ and $R_{q,y} = \Delta_{q,y} - m_q$. Softmax is shift-invariant, so $p_q^\alpha = \mathrm{softmax}(z_q^0 + \alpha R_q)$. Also

$$H_q(\alpha) = \mathrm{LSE}(z_q^0 + \alpha\Delta_q) - \mathrm{LSE}(z_q^0) - \alpha m_q = \log \sum_y p_q^0(y)\, e^{\alpha R_{q,y}}.$$

Thus $H_q(\alpha) = \log \mathbb{E}_{p_q^0}[e^{\alpha R_q}]$. Differentiating gives

$$H_q'(\alpha) = \frac{\mathbb{E}_{p_q^0}[R_q e^{\alpha R_q}]}{\mathbb{E}_{p_q^0}[e^{\alpha R_q}]} = \mathbb{E}_{p_q^\alpha}[R_q] \geq 0,$$

since $R_q \geq 0$, and $H_q''(\alpha) = \mathrm{Var}_{p_q^\alpha}(R_q) \geq 0$. Near $\alpha = 0$, $H_q(\alpha) = \alpha\mu_q + \frac{1}{2}\alpha^2 \mathrm{Var}_{p_q^0}(R_q) + O(\alpha^3)$ (cumulant expansion; not claimed as a new result).

B.3 JENSEN AND SUPPORT BOUNDS

Convexity of $t \mapsto e^{\alpha t}$ for $\alpha \geq 0$ yields $\log \mathbb{E}[e^{\alpha R_q}] \geq \alpha \mathbb{E}[R_q] = \alpha \mu_q$. The bound $0 \leq R_q \leq M_q = \operatorname{range}(\Delta_q)$ yields $H_q(\alpha) \leq \alpha M_q$. Hence $\alpha \mu_q \leq H_q(\alpha) \leq \alpha \operatorname{range}(\Delta_q)$.

B.4 RANGE CLIP BRACKET

Range Clip $\alpha_q^{\text{clip}} = \min(1, h/M_q)$ satisfies $H_q(\alpha_q^{\text{clip}}) \leq h$, so $\alpha_q^{\text{clip}} \leq \alpha_q^\star$. If $\mu_q > 0$ then $H_q(\alpha) \geq \alpha \mu_q$ implies every feasible α obeys $\alpha \leq h/\mu_q$, therefore

$$\min\big(1, h/M_q\big) \leq \alpha_q^\star \leq \min\big(1, h/\mu_q\big).$$

If $\mu_q = 0$, then $\mathbb{E}_{p_q^0}[R_q] = 0$ with $R_q \geq 0$. Softmax has full positive support, so $R_{q,y} = 0$ for every class, $H_q \equiv 0$, and $\alpha_q^\star = 1$.

B.5 STATE STOCHASTIC ORDERING

Fix a common profile R. If p_A first-order stochastically dominates p_B on R (more mass on larger relative displacements), then $e^{\alpha R}$ is increasing in R for $\alpha \geq 0$, so $\mathbb{E}_{p_A}[e^{\alpha R}] \geq \mathbb{E}_{p_B}[e^{\alpha R}]$ and $H_A(\alpha) \geq H_B(\alpha)$. Hence $\alpha_A^\star \leq \alpha_B^\star$ at a common budget h.

B.6 CONFIDENCE INSUFFICIENCY

Let $c \in (\frac{1}{2}, 1)$ and $\varepsilon \in (0, (1-c)/2)$. Set $p_A = (c, 1-c-\varepsilon, \varepsilon)$ and $p_B = (c, \varepsilon, 1-c-\varepsilon)$, both with unique top class 1 and top-1 mass c. For $R = (0, 1, 2)$,

$$H_A(\alpha) = \log\big(c + (1-c-\varepsilon)e^{\alpha} + \varepsilon e^{2\alpha}\big), \qquad H_B(\alpha) = \log\big(c + \varepsilon e^{\alpha} + (1-c-\varepsilon)e^{2\alpha}\big).$$

The difference of the arguments is $(1-c-2\varepsilon)(e^{2\alpha} - e^{\alpha}) > 0$ for $\alpha > 0$, so $H_B(\alpha) > H_A(\alpha)$ and thus $\alpha_B^\star < \alpha_A^\star$ for h strictly between the two curves. For $K > 3$, add $K-3$ extra classes with identical mass $\delta/(K-3)$ and $R = 0$, taking $\delta > 0$ from the residual of classes 2 and 3 while keeping class-1 mass equal to c. Choose δ small enough that $1 - c - 2\varepsilon - \delta > 0$. Both states receive the same extra $R = 0$ mass, so the MGF gap remains $(1-c-2\varepsilon-\delta)(e^{2\alpha} - e^{\alpha}) > 0$.

B.7 SHARED COEFFICIENT

Proposition 1 (Global-coefficient conservatism). *For query q, let $\mathcal{A}_q = \{\alpha \in [0,1] : H_q(\alpha) \leq h\} = [0, \alpha_q^\star]$. A shared coefficient λ satisfying the same pointwise contract for every query in a finite set Q must belong to $\bigcap_{q \in Q} \mathcal{A}_q = [0, \min_{q \in Q} \alpha_q^\star]$. Hence the largest globally safe coefficient is $\lambda_{\text{global}}^\star = \min_{q \in Q} \alpha_q^\star$.*

Feasible sets are prefixes, so a common λ cannot exceed $\min_q \alpha_q^\star$ on a finite query collection Q. This is a statement about admissible retained scale, not realized utility.

B.8 BETA CONTRACTION

H_q is convex, $H_q(0) = 0$, and $H_q(\alpha_q^\star) \leq h$. For $\beta \in [0,1]$, $H_q(\beta \alpha_q^\star) \leq \beta H_q(\alpha_q^\star) \leq \beta h$. Observed $\max H$ may numerically approach βh; that is not an identity.

B.9 SEQUENTIAL COMPOSITION

Let $p_0, \ldots, p_T$ have positive support and $D_\infty(p_{t-1} \| p_t) \leq h_t$. For every label y, $\log(p_0(y)/p_T(y)) = \sum_{t=1}^T \log(p_{t-1}(y)/p_t(y))$. Taking $\max_y$ and using $\max_y \sum_t a_{t,y} \leq \sum_t \max_y a_{t,y}$ yields

$$D_\infty(p_0 \| p_T) \leq \sum_{t=1}^T D_\infty(p_{t-1} \| p_t) \leq \sum_{t=1}^T h_t. \tag{12}$$

This is a HarmCap compositional corollary of a standard max-sum inequality, not a new triangle inequality for D_∞. The experiments evaluate one-step updates only.

On Letter Recognition, two stored queries with the same unique top class and $p_{\max} \approx .62$ can have $\alpha^\star \approx .90$ and $.52$; the full-board controlled residual-mass reassignment in Appendix C supersedes this qualitative example.

C MC80 STATE–UPDATE ALIGNMENT MECHANISM

We reconstruct frozen TALENT-MC80 per-query vectors (p_q^0, Δ_q) on all 80 tasks and 279,852 queries with exact agreement to stored α_{full} and α_{conf} (max absolute difference 0). The diagnostic $M_{\text{conf}} = H_{\text{true}}(\alpha_{\text{conf}}) - H_{\text{proxy}}(\alpha_{\text{conf}})$ uses the same R_q and differs only in state weights. Summaries are task-macro with 10,000 task-bootstrap replicates (seed 20260904). These quantities diagnose the existing log-MGF identity; they are not new headline metrics.

Every true-state violation has $M_{\text{conf}} > 0$; every over-contraction query has $M_{\text{conf}} < 0$. The first-order mean gap $\Delta\mu = \mathbb{E}_{p^0}[R] - \mathbb{E}_{p_{\text{conf}}}[R]$ is positively associated with M_{conf} (task-macro $\rho = .57$ [.48, .66]; 68 positive / 12 negative tasks) but is incomplete relative to the full log-MGF mismatch. Because α_{conf} is chosen at the proxy boundary, the sign of M_{conf} versus E_α is expected from the boundary definition; we retain Table 8 rather than a standalone scatter (task-macro Spearman $\rho = .75$ is expected from the boundary definition, not independent evidence).

A qualitative Letter two-query diagnostic ($\alpha^\star$=.90 vs. .52 under matched confidence) is superseded by the full-board controlled alignment in Table 9; we do not retain a standalone appendix panel.

Residual-mass counterfactuals preserve top-1 class, top-1 probability, K, R_q, the residual-mass multiset, entropy, and K_{eff}, and permute only which non-top class receives which residual mass. Sampling was frozen before outcomes (up to 200 test queries/task, seed 20260906; 15,887 queries). For $K \leq 7$ residual permutations are enumerated when unique; otherwise 100 permutations per query.

Opposite residual alignments (largest residual mass on largest versus smallest R) satisfy $\alpha^\star_{\text{low-}R} - \alpha^\star_{\text{high-}R} \geq 0$ on every task, illustrating Appendix B.5. Predeclared per-bin examples are queries closest to the *task-median* α range, not the maximum.

Within-task normalized-entropy quartiles: true-state violation is 1.1%, 2.8%, 4.4%, 7.5% from Q1 to Q4, but mean $|\alpha_{\text{conf}} - \alpha_{\text{full}}|$ is not monotone (0.36, 0.43, 0.45, 0.42). Neither nominal K nor entropy alone determines the sign of the control error.

Controlled-K synthetic states (Appendix-only) use $c = 0.60$, $K \in \{2, 3, 5, 10, 20, 50, 100\}$, Dirichlet residual concentration $\gamma \in \{0.1, 1, 10\}$, 10,000 replicates per cell, seed 20260907, and R affine-normalized to $[0, 1]$. This family isolates output dimensionality; it is not a real-retrieval result and does not prove that HarmCap becomes more valuable as K grows. Under this generator, mean $|E_\alpha|$ is *not* monotone increasing in K.

Table 7: Frozen TALENT-MC80 state operating points, including the utility and BUR cells underlying the compact main-text ablation. Retention/BUR and violation rates are percentages; U is 100 times CE reduction. Global maxima and task-macro averages of maxima are distinct. Utility is not ranked under a shared safety guarantee.

Controller	Ret.	BUR	Global $H_{\max}$	Mean task $H_{\max}$	Viol.	U [95% CI]
Range Clip	11.75	9.84	0.0932	0.0433	0.00	+0.27 [0.18, 0.38]
Confidence-only	44.90	36.69	0.1646	0.0944	3.94	+0.72 [0.32, 1.15]
HarmCap	85.91	82.37	0.1000	0.1000	0.00	+0.35 [-0.87, 1.69]

Table 8: TALENT-MC80 state-compression diagnostics (task-macro; 10,000 task bootstrap, seed 20260904). M_{conf} and E_α are mechanism diagnostics of the existing log-MGF identity, not new headline metrics.

Statistic	Task-macro	95% CI
mean $\lvert M_{\text{conf}}\rvert$	.066	[.060, .072]
mean $\lvert\alpha_{\text{conf}} - \alpha_{\text{full}}\rvert$	.415	[.366, .464]
under-control rate	3.94%	[2.54, 5.42]
over-control rate	81.5%	[76.2, 86.5]
confidence true-state violation	3.94%	[2.54, 5.42]
$\rho(M_{\text{conf}}, E_\alpha)$	.748	[.706, .786]
$\rho(\Delta\mu, M_{\text{conf}})$	.571	[.482, .657]
$\rho(\Delta\mu, E_\alpha)$	.679	[.621, .732]
$\rho(\lvert M_{\text{conf}}\rvert, \lvert E_\alpha\rvert)$	.711	[.668, .751]

Spearman rows: 80/80 tasks defined; $\rho(M, E)$ and $\rho(|M|, |E|)$ positive on 80/80 tasks; $\rho(\Delta\mu, M)$ 68 positive / 12 negative.

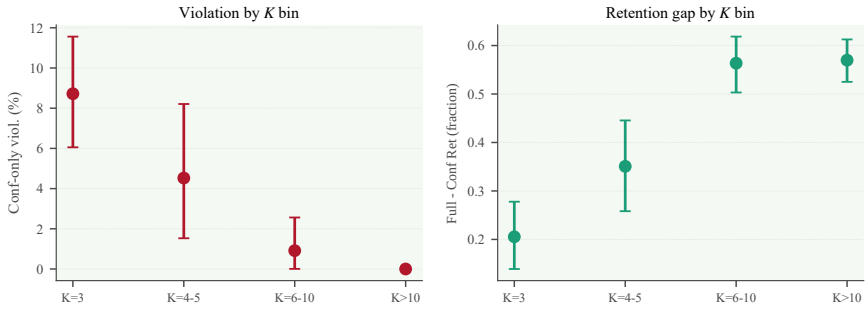

Figure 4: TALENT-MC80 by predeclared K bins (moved from main text). (A) Confidence-only true-state violation (task-bootstrap intervals). Full-state violation is 0 in every bin. (B) Full minus confidence retention. Confidence-only loses exactness on some low/mid-K tasks; in $K>10$ it is exact but more conservative. No monotonic-in-K claim.

Table 9: Real-query residual-mass counterfactual (confidence, entropy, residual multiset, and R fixed). Task-macro within predeclared K bins. Ordered alignment: largest residual mass on smallest R versus largest R.

K bin	n tasks	n queries	mean α range	$>.05$	$>.10$	$>.25$
K=3	25	5,000	.280	.60	.54	.43
K=4–5	16	3,087	.457	.89	.85	.73
K=6–10	27	5,400	.645	.98	.97	.94
$K>10$	12	2,400	.603	1.00	1.00	1.00
Ordered $\alpha^\star$: low-R mass vs. high-R mass (paired difference, task-macro CI)						
K=3	.932 vs. .653; Δ = .280 [.213, .355]					
K=4–5	.876 vs. .419; Δ = .457 [.374, .540]					
K=6–10	.904 vs. .249; Δ = .655 [.600, .705]					
$K>10$	.728 vs. .069; Δ = .659 [.622, .690]					

D ALGORITHMIC AND NUMERICAL VALIDATION

The identities in Appendix B yield the scalar search in Algorithm 1. The uploaded reference core uses float64, 64 iterations, and the search comparison $H \le h$. It returns the feasible lower endpoint; 10^{-8} is the subsequent violation-evaluation tolerance, not slack added inside that search. The zero-range case in Range Clip returns one. Large-vocabulary Newton timings in Appendix J are a distinct implementation stress test, not the execution path of this bisection.

Against high-precision reference roots on 36,407 sampled real queries, the 64-step float64 solver has maximum coefficient error 9.4×10^{-15} and complete reference agreement on the audited contracted queries. Increasing the returned coefficient by 10^{-8} or 10^{-6} crosses the budget

Table 10: Controlled-K synthetic (not a retrieval experiment): $c = 0.60$, 10,000 random-alignment states per cell, seed 20260907. Mean $|E_\alpha|$ is not monotone increasing in K.

K	γ	mean $\lvert E_\alpha \rvert$	under-control rate	mean over-control
2	—	0	0	0
3	0.1 / 1 / 10	.100 / .052 / .015	.50 / .50 / .50	.074 / .036 / .009
5	0.1 / 1 / 10	.087 / .040 / .012	.50 / .50 / .50	.064 / .026 / .007
10	0.1 / 1 / 10	.065 / .023 / .007	.49 / .49 / .49	.046 / .014 / .004
20	0.1 / 1 / 10	.043 / .014 / .005	.50 / .50 / .50	.028 / .008 / .002
50	0.1 / 1 / 10	.026 / .008 / .003	.49 / .50 / .50	.015 / .004 / .001
100	0.1 / 1 / 10	.018 / .006 / .002	.50 / .51 / .50	.010 / .003 / .001

Table 11: Equation-to-implementation mapping for the supplied code snapshot. This mapping identifies inspected code, not a claim that every historical experiment can be rerun from the small review package.

Operation	Source file	Function
Proposed direction	`core.py`	`build_update`
Exact consequence	`consequence.py`	`exact_h`
Maximal ray coefficient	`projection.py`	`harmcap_alpha_star`
Controlled output	`core.py`	`apply_harmcap`
Range envelope	`range_clip.py`	`alpha_range_clip`
Retention / BUR	`harmcap_ops_SOURCE.py`	`retained_scale` / `bur`

boundary in every eligible case. The p99 coefficient error decreases from 3.8×10^{-3} at $B = 8$ to 4.8×10^{-15} at $B = 64$. Pure float32 is not boundary-equivalent: on 6,000 audited queries it yields 788 false-safe and 1,489 false-unsafe decisions, including 1,163 outputs that violate the budget under float64 re-evaluation. Formal experiments therefore use float64 with the declared 10^{-8} evaluation tolerance rather than claiming strict floating-point certification.

A 154,840-case synthetic stress suite covers extreme logits, near-zero and constant-shift updates, sparse/spiked directions, and large class supports. It retains 58 approximately-10^{-8} numerical boundary failures for audit. The ordering $\alpha^{\text{clip}} \le \alpha^\star$ holds on every stored real suite, and the label-aware closed-form full-set oracle agrees with constrained optimization to within approximately 2×10^{-15}. The oracle is evaluation-only and is used to quantify design headroom.

E Reproducibility and Protocol Boundaries

The release preserves the frozen prediction pairs, split metadata, controller outputs, and integrity hashes used for every reported result. Protocols are never pooled when they differ in predictor family, retrieval construction, support definition, split, or export unit. Table 12 states the role of each suite. TALENT intervals use 10,000 task-bootstrap replicates with seed 20260904, and all exact-harm checks use complete finite class support with evaluation tolerance 10^{-8}. The final fair-baseline round uses 143 frozen pair files and does not tune budgets, task sets, or controller definitions on test labels. The shared-gate selector audit uses two predeclared label-free masks applied identically to all controllers; historical nested UtilityFloor artifacts are retained only as prior supporting evidence and are not claimed as reproduced in this round. The published TabR alignment reports paper numbers separately from our 15-seed protocol-aligned reruns and does not perform significance tests against published means.

TabR-36 uses a local reimplementation of the published TabR architecture: both f_0 and f_{ret} share one locally trained checkpoint, CPU FAISS ℓ_2 with k=16, and d_{main}=64. TALENT-120 is a train-only logistic classifier with cosine top-K=20 retrieval, top-1 blend with fixed retrieval coefficient 0.25, clip $[10^{-7}, 1 - 10^{-7}]$, and logits $z = (0, \text{logit}(p))$ recomputed on official splits. The frozen nested export still lacks a recoverable architecture. The independent UCI and RAUS protocols use a logistic-regression base and kNN-conditioned proposals.

Published predictor rows are descriptive substrate references rather than controller comparisons. TabICL completes 120/120 TALENT tasks. TabPFN completes 93/120; the 27 omitted tasks exceed its stated size constraints and are exclusions rather than failed or imputed predictions. CatBoost supplies no matched retrieval/no-retrieval path in the release. Building such a path would require specifying and training a new retrieval wrapper, so its predictive score cannot be substituted for a HarmCap experiment.

E.1 Metric Definitions and Interpretation

For a query with revealed label y_q, define its utility after applying coefficient a as

$$u_q(a) = \ell_{y_q}(z_q^0) - \ell_{y_q}(z_q^0 + a\Delta_q).$$

The released utility helper uses this cross-entropy reduction. Reported U multiplies the task-macro mean by 100; the historical "pp" label is a reporting convention and does not mean accuracy percentage points. For each task, let $B = \{q : u_q(1) > 10^{-12}\}$. The released beneficial-update retention is

$$\text{BUR} = \frac{\sum_{q \in B} \max(u_q(a_q), 0)}{\sum_{q \in B} u_q(1)}.$$

It is undefined when the denominator is zero and is then excluded from the finite-BUR average with its count reported, not imputed as zero. BUR is an outcome-based, positive-utility ratio, not a fraction of correctly classified instances and not necessarily bounded by one. It cannot be used at inference without labels.

Retention follows Eq. (11), using the Euclidean norm and denominator floor 10^{-12} in the inspected helper. For exact same-ray nonzero updates above this floor it equals the applied coefficient. For an identically zero proposal the recorded ratio is zero even though the largest admissible coefficient is one. These conventions explain why an unconstrained row can have average retention slightly below 100%.

NLL evaluates the probability assigned to the true class; Brier score evaluates squared probability error; balanced accuracy is the average class recall (Gneiting & Raftery, 2007; Brodersen et al., 2010). The external score differences retain the normalization and clipping convention of their supplied exports. The small review repository does not contain the external metric runner, so a different Brier normalization is not retrospectively imposed on the reported values. Balanced-accuracy differences in the main table are converted from proportions to percentage points.

Task-bootstrap intervals describe uncertainty in the mean across evaluated tasks. Wilcoxon statistics concern a different paired-distribution target; a mean-difference CI and a Wilcoxon test need not give the same conclusion. Neither an interval crossing zero nor a nonsignificant test proves equivalence. The closure report additionally used engineering tolerances ($+0.02$ NLL, $+0.01$ Brier, and -0.01 balanced-accuracy proportion), but these are not treated as a formal non-inferiority guarantee in the manuscript. Counts, rates, and averages of task maxima are kept distinct.

Table 12: Evaluation suites. Different rows are different protocols and are never pooled.

Suite	Unit	Role
TabR-36	103,367 queries; 36 tasks	Dual path (local TabR architecture; frozen-manifest subsample)
TALENT-120	315,600 queries; 120 tasks	Binary breadth replay (logistic–kNN blend)
TALENT-MC80	279,852 queries; 80 tasks; K=3–100	Formal multiclass mechanism/controller board (logistic–cosine-kNN)
OpenML-CC18	87,494 queries; 72 tasks	External public benchmark replication; 72/72 formal paired tasks
TabZilla-36	201,934 queries; 35 formal tasks	Hard-suite replication; 1 pre-result OOM exclusion (`albert`)
TabICL external subset	16 tasks	Second author-constructed kNN-context substrate; pre-stratified before outcomes
Native TabR fair board	16 CC18 tasks	LEVEL-B same-checkpoint retrieval ablation; strong safe-controller comparison
TabR original benchmark	CH/AD/OT/HI; 15 predictive seeds	LEVEL-P2 published alignment plus same-checkpoint HarmCap overlay
Shared-gate audit	Native TabR16, TabICL16, TabR-original4	Same predeclared label-free accept masks for all controllers
TabDPT / TabPFN-kNN / LoCalPFN	supporting frozen boards	Older interface evidence; not pooled with the final fair-controller boards

F TabR-36 frozen selection summary

TabR-36 is a pre-specified stratified subsample of 36 of the 120 binary TALENT tasks (seed 20260830), selected on size, dimension, categorical mix, imbalance, and missingness before local TabR training or controller outcomes. All 36 selected tasks completed; the remaining 84 binary tasks were not attempted on this compute-limited local dual-path board and were not excluded after inspecting HarmCap–Range Clip differences. TALENT-MC80 is a separate all-80 multiclass board and is never pooled with TabR-36.
The frozen manifest hashes are `subset_meta.json`: `5b4cf7be...775928b7bcb071b0` and `subset_selection.csv`: `2a3dd2a9...d1fb3d3b78994db6`. This identifies the exact local-board selection without reproducing several pages of task names in the manuscript.

G Predictive breadth

Standalone TALENT official-split ranks check whether the tabular board is weak or CatBoost-dominated. Rank, NLL, and accuracy of the ten TALENT predictors below use the same 93-task intersection. Table 15 adds dual-path ICL retrieval operating points ranked against those ten predictors on the available 93-overlap; they are not official checkpoints and do not rewrite the ten-model rank. Figure 6 reports pairwise win–loss counts versus CatBoost on each official method's common tasks.

H External tabular validation

Published TabR anchor. For the four completed classification datasets from TabR's original benchmark, the protocol-aligned Stage-A rerun uses the published fixed split, Accuracy metric, context size $m = 96$, and published/best configuration for 15 seeds without retuning. The published/ours means are CH .862/.8626, AD .870/.8700, OT .825/.8247, and HI .729/.7301, giving absolute gaps of .000600, .000008, .000259, and .001100, respectively. This supports a LEVEL-P2 protocol-aligned comparison, not an exact LEVEL-P1 reproduction of the full hyperparameter-search procedure. Covertype was still running at the artifact freeze and is not included in the current claim. The HarmCap overlay uses same-checkpoint matched prediction pairs and is evaluated separately from these 15-seed predictive means.

Task accounting. The final external-validation accounting uses per-task result rows and matched-pair archives as the source of truth. OpenML-CC18 has 72/72 registered/formal paired tasks with no exclusions. TabZilla has 36 registered tasks and 35 formal paired tasks; the single exclusion is `albert` (OpenML task 189356), which fails with `MODEL_OOM` before any controller output is recorded. No task is excluded based on HarmCap retention, utility, or violation outcomes.

CC18 controller-family replication. A 12-task CC18 subset was pre-stratified before controller outcomes. Table 17 reports task-macro retention and test exact-H violations. In this subset, Range Clip and HarmCap are the only zero-violation controllers; HarmCap is substantially less conservative.

Second retrieval-conditioned substrate. The second-substrate experiment uses a pre-stratified 16-task TabICL kNN-context interface. It is an author-constructed context path rather than an official TabICL retrieval method. Range Clip retains .1188 of the proposal and HarmCap .4791, for a paired ΔRetention of .3604 [.2386, .4925] with 16/0/0 wins and zero HarmCap violations. The paired standard-metric changes are ΔNLL $= +.0034$, ΔBrier $= +.00012$, and Δbalanced-accuracy $= -.00014$.

Standard-metric confidence intervals. Table 18 reports the complete task-bootstrap intervals used in the main text. The closure report's engineering thresholds flag small aggregate changes on both suites. This is an interpretation of the supplied intervals, not a formal non-inferiority or distribution-free guarantee. Per-task NLL differences remain heterogeneous.

Dataset difficulty. Using base NLL as a pre-controller task difficulty measure, the Spearman correlation between base NLL and the HarmCap–Range Clip retention gain is negative on both suites: $\rho = -.447$ ($p = 8.4 \times 10^{-5}$, $n = 72$) on CC18 and $\rho = -.314$ ($p = .066$, $n = 35$) on TabZilla. We therefore do not claim that harder datasets systematically produce larger HarmCap gains.

Available protocol evidence. The closure records report task counts, query counts, subset sizes, and aggregate outputs. The supplied files do not include the external repeat/fold indices, full 12- and 16-task manifests, cross-suite duplicate-ID map, the exact second-TabICL context-construction runner, or the complete Brier-score implementation. Accordingly, the reported numbers describe the exported task evaluations, not a verified execution of every official cross-validation repeat or a dataset-disjoint external population. These materials are needed for full experiment-level reproduction; re-rendering the summary tables alone does not provide that reproduction.

Native TabR closure. A separate 16-task CC18 subset uses the published TabR architecture with official TabR-S-D hyperparameters and CPU FAISS. The matched pair is constructed within one checkpoint: the retrieval-enabled path is the standard forward pass, while the base path sets the retrieval contribution `context_x` to zero before the shared downstream blocks and head. This is a LEVEL-B same-checkpoint ablation, not an official retrieval-off API. Range Clip retains 15.3% and HarmCap 21.1%; the paired gain is +5.79 pp [3.28, 8.82], with 16/0/0 task-level wins and zero HarmCap violations. Raw TabR records 4,020 exact-H violations on this subset, so predictive quality and contract compliance are reported separately.

Table 13: Frozen TabR-36 selection audit. The full task manifest is released with the code package; the paper keeps only the pre-result accounting needed for review.

Item	Count	Audit note
Binary TALENT registry	120	official-split binary tasks
Selected local TabR board	36	frozen stratified manifest, seed 20260830
Completed selected tasks	36	no post-result exclusions
Not attempted on local board	84	compute-budget only; remain in TALENT-120
Non-binary TALENT tasks	180	excluded by task type before outcomes

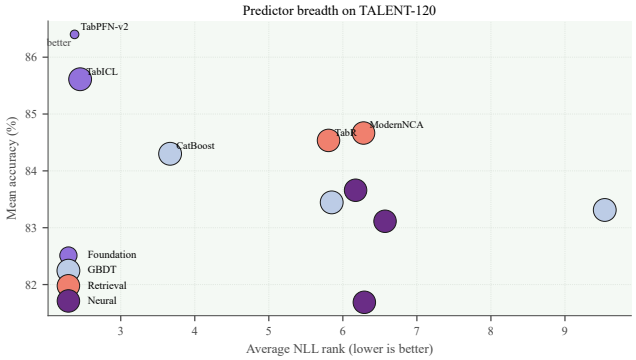

Figure 5: TALENT official-split predictors: 93-task accuracy versus NLL rank.

Table 14: Standalone TALENT official-split predictors. Rank, accuracy, and NLL are task-macro on the 93-task TabPFN intersection. The $n/120$ column is completed-task coverage, not the intersection size. The TabR row is the standalone TALENT predictor, not a matched retrieval-control result. Dual-path ICL ranks: Table 15.

Predictor	Venue	$n/120$	Rank↓	Acc. (%)	NLL↓	ΔNLL vs CB
TabPFN-v2 (Hollmann et al., 2025)	Nature'25	93	2.38	86.4	.293	−.026
TabICL (Qu et al., 2025)	ICML'25	120	2.45	86.1	.297	−.022
CatBoost (Prokhorenkova et al., 2018)	NeurIPS'18	120	3.67	84.8	.319	.000
TabR (TALENT standalone) (Gorishniy et al., 2024)	ICLR'24	119	5.81	84.6	.343	+.024
XGBoost (Chen & Guestrin, 2016)	KDD'16	120	5.85	83.9	.339	+.020
ResNet (Gorishniy et al., 2021)	NeurIPS'21	119	6.17	84.0	.353	+.034
ModernNCA (Ye et al., 2025)	ICLR'25	119	6.28	84.9	.409	+.090
TabM (Gorishniy et al., 2025)	ICLR'25	119	6.29	82.0	.359	+.040
MLP (Gorishniy et al., 2021)	NeurIPS'21	119	6.57	83.4	.353	+.034
LightGBM (Ke et al., 2017)	NeurIPS'17	120	9.54	83.7	.954	+.635

Table 15: Dual-path ICL retrieval path, ranked vs the ten TALENT official-split predictors on available 93-overlap (TabPFN-kNN / LoCalPFN: $89/93$). Not a controller table and not pooled with Table 14. The TabICL+kNN wrapper is omitted.

Predictor	Venue	$n/120$	Rank↓	Acc. (%)	NLL↓	ΔNLL vs CB
TabDPT (Ma et al., 2025)	NeurIPS'25	120	3.18	85.7	.300	−.019
LoCalPFN (Thomas et al., 2024)	NeurIPS'24	116	4.99	85.3	.310	−.006
TabPFN-kNN (Thomas et al., 2024)	NeurIPS'24	116	5.36	84.4	.321	+.005

Table 16: External benchmark task accounting. "Formal" denotes complete matched base/retrieval pairs and both Range Clip and HarmCap controller outputs.

Suite	Registered	Predictor	Pair	Formal	Exclusion
OpenML-CC18	72	72	72	72	none
TabZilla-36	36	35	35	35	1 × `MODEL_OOM`

External budget robustness. Replaying the frozen CC18 and TabZilla matched pairs at $h \in \{.01, .025, .05, .1, .2, .5\}$ gives positive HarmCap–Range Clip Retention differences with task-bootstrap CIs excluding zero at every budget, while HarmCap records zero exact-H violations throughout. This supports robustness of the safety–efficiency ordering across the tested budgets; it is not a utility-robustness or universal-threshold claim.

Deferred and unavailable components. Temporal-shift experiments are deferred and no shift-robustness claim is made. TabPFN is license-gated in the current environment and TabDPT is blocked by the Windows/Triton environment for the new external runs. The native TabR closure uses the same-checkpoint LEVEL-B ablation described above; an official LEVEL-A retrieval-off API remains unavailable. None of these interface limitations is imputed.

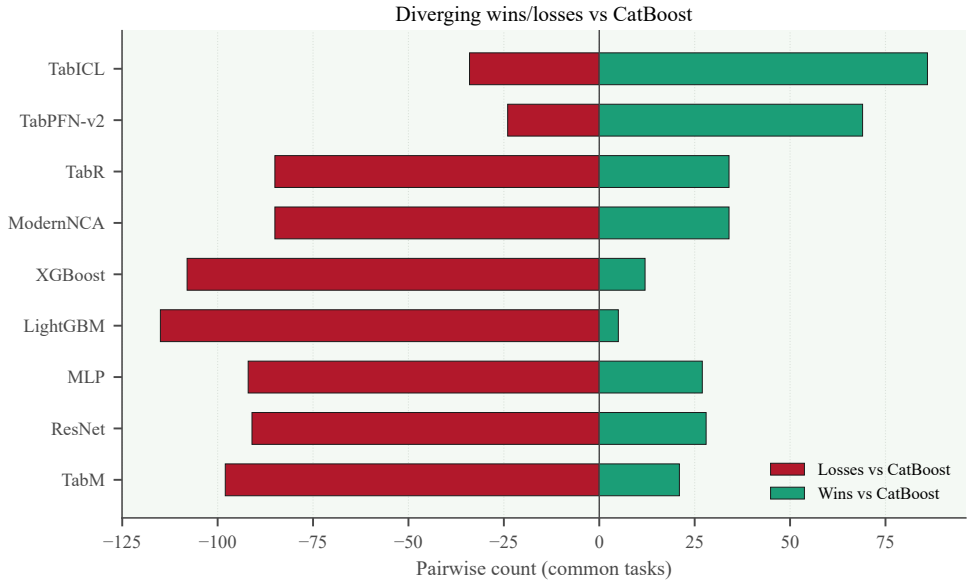

Figure 6: Pairwise substrate comparisons versus CatBoost on TALENT-120. Green: wins; salmon: losses. Not controller comparisons.

Table 17: CC18 external controller-family replication on a 12-task pre-stratified subset. Parameters for shared/norm/divergence controllers are selected on development data and frozen before test.

Controller	Tasks	Ret. mean	95% CI	Max-H macro	Test viol.
Global	12	.465	[.430,.508]	.0941	7
Centered L_2	12	.210	[.124,.310]	.0881	2
Centered L_∞	12	.197	[.116,.291]	.0833	1
Range Clip	12	.182	[.095,.279]	.0673	**0**
FKL	12	.414	[.366,.467]	.0887	2
RKL	12	.394	[.343,.452]	.0902	3
Rényi-2	12	.442	[.397,.492]	.0893	7
Rényi-4	12	.483	[.435,.531]	.0898	4
Rényi-8	12	.568	[.511,.622]	.0898	5
Rényi-16	12	.699	[.621,.774]	.0884	6
HarmCap	12	**.947**	**[.921,.969]**	.1000	**0**

Table 18: Standard predictive metrics for HarmCap minus Range Clip on the external suites.

Suite	ΔNLL [95% CI]	ΔBrier [95% CI]	ΔBal. Acc. [95% CI]
OpenML-CC18	−.0137 [−.0423, .0063]	−.0030 [−.0058, −.0006]	+.0001 [−.0023, .0025]
TabZilla-36	−.0209 [−.0740, .0103]	−.0037 [−.0084, .0002]	+.0024 [−.0016, .0068]

Fair safe-controller board. The final fair-baseline audit adds two stronger label-free zero-violation alternatives to Range Clip. Moment Envelope stays on the realized logit ray and uses the sufficient bound

$$H(\alpha) \le \log\left[1 + \frac{\mu}{M}\left(e^{\alpha M} - 1\right)\right], \qquad M = \max R,\ \mu = \langle p^0, R\rangle,$$

then takes the largest feasible α. Probability-Mix-Max instead follows $p_\beta = (1-\beta)p^0 + \beta p^r$ and uses the closed-form classwise constraint

$$\beta \le \frac{(1-e^{-h})p^0_y}{p^0_y - p^r_y} \quad \text{for each } y \text{ with } p^r_y < p^0_y,$$

with $\beta = 1$ when no class decreases. It enforces the same probability lower-bound form of the D_∞ contract but is maximal in probability space rather than on the realized logit ray. Consequently, logit norm-ratio Retention is not used to compare Probability-Mix-Max with the same-ray controllers.

HarmCap has the largest same-ray Retention, as expected from exact maximality on the realized logit ray. Predictive NLL follows a different ordering across substrates, which is why the main paper separates admissibility and proposal preservation from predictive utility.

I Additional retrieval-interface, fair-controller, and risk-object evidence

Additional retrieval interfaces. Earlier frozen supporting boards using TabDPT, TabPFN-kNN, LoCalPFN, a local TabR board, and an author-constructed TabICL wrapper show the same qualitative HarmCap–Range Clip retention ordering under their documented pair constructions. We do not pool these operating points with the external benchmark results because the interfaces are not identical: TabDPT

Table 19: Cross-suite fair safe-controller audit at $h = .1$. NLL is task-macro; all displayed safe-controller violation rates are zero. Retention is shown only for same-ray controllers because probability-space interpolation is off the realized logit ray.

	CC18		TabZilla		TabICL16		Native TabR16		TabR CH/AD/OT/HI	
Controller	Ret.	NLL	Ret.	NLL	Ret.	NLL	Ret.	NLL	Ret.	NLL
Range Clip	.262	.446	.352	.524	.119	**.289**	.153	.579	.142	.831
Moment Envelope	.756	**.426**	.814	**.497**	.468	.292	.210	.521	.277	.779
Prob.-Mix-Max	–	.434	–	.504	–	.293	–	**.503**	–	**.754**
HarmCap	**.923**	.433	**.907**	.503	**.479**	.292	**.211**	.518	**.278**	.778

and TabPFN-kNN use same-checkpoint context changes, LoCalPFN is cross-model because local fine-tuning changes model state, and the TabICL+kNN path is our wrapper rather than an official retrieval interface. Complete operating points remain in the released experiment artifacts; the main paper instead emphasizes the protocol-aligned TabR anchor, the external TabICL16 interface, and the native same-checkpoint TabR closure.

Different risk objects. Forward/reverse KL and adapted conformal-risk controls are useful reference points but do not certify the same object as HarmCap. A factorial audit confirms that satisfying a population- or selection-level control does not imply the per-update exact-H contract; composing such a selector with Range Clip or HarmCap can enforce the latter without changing the selector's own interpretation. We therefore do not rank these methods as if they shared a common guarantee. Exact-H violations under our audit should not be read as failures of the comparators' native risk guarantees.

Shared-gate selector audit. The final selector check does not reuse the historical UtilityFloor board as a reproduced result. Instead, two label-free gates are fixed in advance and the same accept mask is applied to Base, Raw, Range Clip, HarmCap, Moment Envelope, and Probability-Mix-Max on each frozen board; rejected queries fall back to Base. On Native TabR, the confidence-delta gate yields HarmCap–Range Clip $\Delta\text{NLL} = -0.0308$ $[-0.0565, -0.0126]$, whereas on TabICL16 the corresponding contrast is approximately zero. These results support modular composition of admissibility control and usefulness selection rather than a universal ranking after gating.

J Budget sensitivity and computational scope

The headline operating point is $h = .1$. The classwise interpretation $p^{\text{ctrl}}(y) \geq e^{-h} p^0(y)$ gives the following probability floors for the pre-specified sweep: 99.0% ($h = .01$), 97.5% (.025), 95.1% (.05), 90.5% (.1), 81.9% (.2), and 60.7% (.5). Earlier frozen CC18 and TabZilla replays preserve the HarmCap–Range Clip Retention ordering throughout this grid. The final fair-baseline package additionally evaluates Native TabR and TabICL16 at all six budgets. All safe controllers remain zero-violation there, but predictive ordering is not stable across geometries: Probability-Mix-Max has lower NLL than HarmCap at every tested Native-TabR budget, whereas on TabICL16 Range Clip is often lowest-NLL despite being the most conservative same-ray controller. Thus the sweep supports robustness of the contract and exposes a geometry-dependent utility trade-off rather than a single dominant operating rule.

The runtime comparison concerns only the controller. In applications where retrieval or the predictor dominates latency, end-to-end differences may be much smaller; conversely, the current 64-step implementation is not optimized for speed. The iteration audit shows that substantially fewer bisection steps approximate the reference root closely, but the formal experiments retain 64 steps for consistency.

K Limitations

HarmCap bounds a relative worst-label cross-entropy increase for finite-class softmax outputs. It does not certify absolute correctness, calibration, fairness, or realized utility, and a wrong base prediction can remain wrong. For a generic finite-label loss one can define an analogous worst-label consequence, but the monotonicity, log-MGF representation, and efficient one-dimensional solution used here need not hold for one-vs-rest, margin-based, regression, structured, or free-running generative outputs.

The controller is path-specific. HarmCap is exact and maximal on the realized logit-space ray, but it cannot correct a poor proposal direction. The label-aware oracle decomposition shows substantial headroom outside that ray, and the fair-controller audit demonstrates the same limitation without labels: a maximal probability-space interpolation can yield lower NLL than HarmCap on some substrates while satisfying the same classwise probability lower bound. Thus choosing the admissible geometry is itself a modeling decision; HarmCap should not be interpreted as the utility-optimal element of the full feasible set.

Predictive usefulness remains separate from admissibility. The shared-gate audit applies identical predeclared label-free masks to every controller and shows that gating can change predictive outcomes without changing the underlying consequence contract. These gates are intentionally not integrated into HarmCap, and they provide no distribution-free usefulness guarantee under deployment shift. Historical UtilityFloor artifacts are retained as prior supporting evidence but are not claimed as reproduced by the final fair-baseline round.

The numerical implementation is also finite-precision. Float64 bisection agrees with high-precision reference roots to approximately machine precision on the real-query audit, but pure float32 yields false-safe near-boundary decisions. The synthetic stress audit retains 58 approximately-10^{-8} boundary excesses, so we do not claim strict floating-point certification. Formal results use float64 and the declared 10^{-8} evaluation tolerance. The 64-step solver is also materially slower than the closed-form safe alternatives in a controller-only CPU audit, leaving adaptive precision and faster root solvers as engineering opportunities.

External evidence spans logistic–kNN proposals on CC18 and TabZilla, an author-constructed TabICL context interface, a native same-checkpoint TabR ablation, and protocol-aligned experiments on four datasets from TabR's original classification benchmark. This substantially reduces dependence on one proposal family but does not establish universality across learned retrievers or production retrieval systems; the largest absolute Retention gains still occur on the logistic–kNN boards. The TabR base path is a LEVEL-B same-checkpoint ablation obtained by zeroing `context_x`, not an official retrieval-off API. Broader learned-retriever and temporal-shift studies remain future work.

Finally, h is externally specified and application-dependent. Its multiplicative probability-floor interpretation makes the operating point interpretable, but it does not determine an application-specific tolerance by itself. NLL, Brier, and balanced-accuracy effects can be heterogeneous even when the exact-H contract is satisfied, and the negative hardness correlations observed on CC18/TabZilla do not support a monotone difficulty claim. All retained evidence is therefore reported at its original protocol scope.

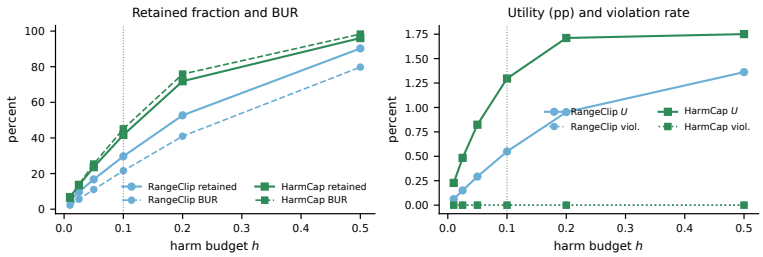

Figure 7: Harm-budget sensitivity on the TALENT-120 replay. The vertical marker denotes the main operating point $h = .1$. This replay illustrates the operating scale of h and is not pooled with the fair-controller boards.

Table 20: Controller-only CPU latency on the common synthetic runtime setting ($n = 1024$, $K = 5$, float64; 30 repeats). Timings exclude retrieval construction and predictor inference.

Controller	Median μs/query	Relative to HarmCap
Range Clip	0.270	106× faster
Moment Envelope	0.147	195× faster
Probability-Mix-Max	0.196	146× faster
HarmCap (64-step bisection)	28.660	1.0×