

# Curated to Win: Certifying Reward-Model and Judge Selection Against Evaluation-Set Curation

Adrian Erlikhman

Los Angeles Community College District  
erlikhman.adrian@gmail.com

## Abstract

Automated research and self-improvement pipelines select a reward model or LLM judge by its accuracy on a validation set, and that set is often curated by the pipeline or its designers. We propose a *curation certificate* for this selection: how much of the validation set could be removed without changing the winner. For weighted-accuracy benchmarks the certificate has a closed form. Measured in benchmark weight it is at least the gap between the top two models, and measured in items it is computed exactly by sorting per-item score differences. Linear programs over item weights, rounded to verified item subsets, show that the weight bound is tight near the top of RewardBench (118 models, 2985 items) and RewardBench 2 (188 models, 1753 items). RewardBench’s leader is certified against the removal of any 11 items, and removing 12 particular items makes the runner-up the winner; RewardBench 2’s leader is certified against any 42. At a curation budget of 1% of benchmark weight, 3 RewardBench models could win. The certificate also works on small validation sets: with 200 random items, selections whose top-two gap exceeds 2 points agree with the full-benchmark winner 87% of the time, against 43% for the rest. We give a selection protocol that reports the certificate, or the set of possible winners when the certificate is below the curation budget.

## 1 Introduction

Systems that automate research or improve themselves need an evaluator: a reward model to optimise against, or an LLM judge that ranks candidate outputs (Lu et al., 2024; Zheng et al., 2023). Evaluators are chosen by meta-evaluation, that is, by their agreement with human preference labels on a benchmark such as RewardBench (Lambert et al., 2024; Malik et al., 2025). In an automated pipeline the items used for this choice are often selected by the pipeline itself: a filtered subset, a

domain slice, or a small validation set. If the ranking of evaluators depends on which items are kept, whoever selects the items also selects the evaluator, which is reward hacking one level above the reward (Gao et al., 2022). A selection should therefore come with a statement of how much curation it survives.

Leaderboards are known to be sensitive to benchmark choice, answer format and private testing (Dehghani et al., 2021; Alzahrani et al., 2024; Singh et al., 2025; Zhang and Hardt, 2024), and to targeted edits of individual items (Dubrovsky et al., 2025). Gordienko et al. (2026) count the *datasets* a developer must train on to top a multi-task leaderboard and show that the problem is NP-hard under mean win rate, and Oyarhoseini et al. (2026) show with influence approximations that sub-1% edits to pairwise matches can change the leader of a Bradley–Terry arena. Both describe how rankings can be attacked. We instead give the party selecting an evaluator a certificate it can compute and report, for the weighted-accuracy scores used in reward-model and judge meta-evaluation. Our contributions are:

- Two curation certificates for the winner of a weighted-accuracy benchmark: a lower bound in benchmark weight and an exact value in items (Props. 1–2).
- Evidence from linear programs on RewardBench and RewardBench 2 that the weight bound is tight near the top of both leaderboards.
- Two applications: sets of possible winners at a given curation budget, and certificates computed on small validation sets, which separate reliable selections from unreliable ones.

## 2 Setup

A benchmark has items  $i = 1, \dots, N$  with official weights  $w_i^0 \geq 0$ ,  $\sum_i w_i^0 = 1$ . Model  $m$

has per-item score  $r_{mi} \in [0, 1]$  and official score  $S_m = \sum_i w_i^0 r_{mi}$ . A *curation* keeps a set  $K$  with kept share  $f = \sum_{i \in K} w_i^0$  and rescores every model on it,  $S_m^K = \sum_{i \in K} w_i^0 r_{mi} / f$ . We call  $1 - f$  the *removed share*;  $K$  *crowns*  $m$  if  $S_m^K > S_k^K$  for every  $k \neq m$ .

**Data.** RewardBench v1 publishes per-item results for 118 models on 2985 preference pairs in 23 subsets (Lambert et al., 2024). We rebuild each official score from them: a section score is the count-weighted mean of its subset accuracies, with math-prm upweighted as in the released code, and the overall score is the mean of the four section scores. For four models, from the leader to a 0.5B chat model, our subset accuracies match RewardBench’s published values exactly. Result files are aligned by position; a text check found no misaligned items. RewardBench 2 (Malik et al., 2025) publishes results by item id. We use the 188 models with results on all 1753 items of its five accuracy subsets, scored as the mean of subset accuracies, and exclude the Ties subset, which uses a different metric.

### 3 Curation Certificates

Let  $a$  be the winner and, for each challenger  $c$ , let  $g_c = S_a - S_c > 0$  and  $d_i^c = w_i^0(r_{ai} - r_{ci})$ .

**Proposition 1.** *Every  $K$  with  $S_c^K \geq S_a^K$  removes a share of at least  $g_c$ . Hence the winner is certified against any curation that removes a share below  $g^* = \min_c g_c$ , the gap to the runner-up.*

*Proof.*  $f(S_a^K - S_c^K) = \sum_{i \in K} d_i^c = g_c - \sum_{i \notin K} d_i^c \geq g_c - (1 - f)$ , since  $d_i^c \leq w_i^0$ ; this is positive whenever  $1 - f < g_c$ .  $\square$

**Proposition 2.** *Sort  $d_{(1)}^c \geq d_{(2)}^c \geq \dots$  and let  $k_c = \min\{k : \sum_{j \leq k} d_{(j)}^c > g_c\}$ . Removing a set  $R$  makes  $c$  overtake  $a$  if and only if  $\sum_{i \in R} d_i^c > g_c$ , so no removal of fewer than  $k_c$  items does, and removing the top  $k_c$  items does. The winner is therefore certified against the removal of any  $k^* - 1$  items, where  $k^* = \min_c k_c$ , and this is exact.*

*Proof.* From the identity above,  $S_c^K > S_a^K$  iff  $\sum_{i \notin K} d_i^c > g_c$ . Any  $k$  removed items contribute at most  $\sum_{j \leq k} d_{(j)}^c$ .  $\square$

Both certificates need only the per-item score table:  $g^*$  is read off the leaderboard, and  $k^*$  costs one sort per challenger. Prop. 1 is a lower bound, since removing items that favour the winner over

one challenger can let a third model overtake both. To test how tight it is, we compute for each model  $m$  the largest kept share  $f$  at which some curation crowns it. For fixed  $f$ , relaxing the subset to weights  $w$  gives a linear program:

$$\begin{aligned} \max_{w, t} \quad & t \quad \text{s.t.} \quad \sum_i w_i(r_{mi} - r_{ki}) \geq t \quad \forall k \neq m, \\ & 0 \leq w_i \leq w_i^0 / f, \quad \sum_i w_i = 1, \end{aligned}$$

and  $m$  is crownable at  $f$  if  $t^* > 0$ . We scan  $f$  downward from 1 and bisect, round each solution to an item subset by keeping the items at or near their upper bound, and accept it only if  $m$  is still strictly first after rescoring, so every reported cost belongs to a verified subset. A *section-preserving* variant also fixes each section’s total weight, so the curated set keeps the benchmark’s composition; since any verified section-preserving subset is also valid without the constraint, we report the unconstrained cost as the smaller of the two. For ordinal aggregates such as mean win rate, the analogous selection problem is NP-hard (Gordienko et al., 2026); for weighted accuracy each check is a linear program.

**Possible winners.** At curation budget  $\delta$ , let  $P(\delta)$  be the set of models that some curation removing at most  $\delta$  can crown. By Prop. 1,  $P(\delta) \subseteq \{m : S_a - S_m \leq \delta\}$ , and every model with a verified LP cost at most  $\delta$  is in  $P(\delta)$ , which brackets its size.

### 4 Results

**Certificates of the two leaders.** RewardBench’s winner leads by  $g^* = 0.67$  points, so it survives any curation removing less than 0.67% of benchmark weight; the smallest verified crowning curation removes 0.68%, so the certificate is tight to within 0.01 points. In items, the winner survives the removal of any 11 items, and removing 12 particular items makes the runner-up the new winner, which we verified by rescoring. (The LP, which minimises weight rather than count, used 19 items.) RewardBench 2’s winner leads by 5.22 points, needs 5.30% removed, and is certified against any 42 items.

**The weight bound is tight near the top.** Every model on both benchmarks can be crowned. The ratio of removed share to gap has a median of 1.00 on RewardBench and at most 1.02 in its top 20 (Figure 1); the largest ratio over all models is 2.00. Every top-20 model can be crowned

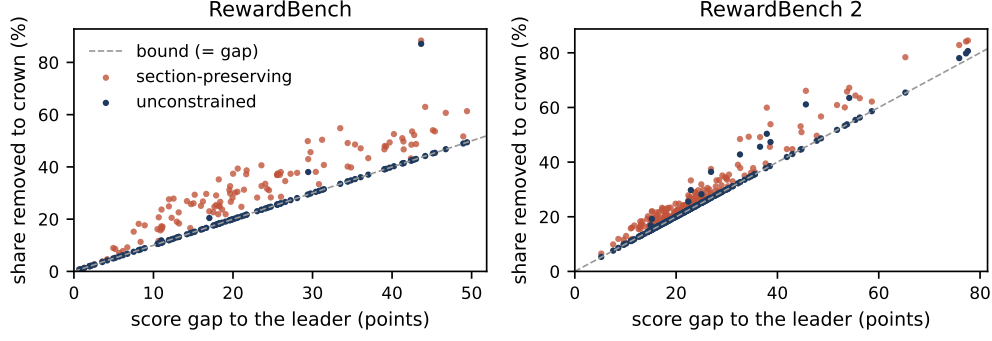


Figure 1: Share of the benchmark removed to crown each model, against its official gap to the winner. Every point lies on or above the dashed line  $y = x$ , the bound of Prop. 1. Unconstrained curation reaches the bound for nearly every model; keeping each section’s share costs more, at most 2.31 times the gap.

| Budget $\delta$ | 0.5% | 1% | 2%  | 5% | 10% |
|-----------------|------|----|-----|----|-----|
| RewardBench     | 1    | 3  | 5–6 | 11 | 19  |
| RewardBench 2   | 1    | 1  | 1   | 1  | 6   |

Table 1: Number of possible winners  $|P(\delta)|$  at curation budget  $\delta$  (share of benchmark weight). A range gives the lower bound from verified LP curations and the upper bound from Prop. 1.

| $n$ | RewardBench |         |     | RewardBench 2 |          |     |
|-----|-------------|---------|-----|---------------|----------|-----|
|     | all         | cert.   | not | all           | cert.    | not |
| 100 | 35          | 64 (15) | 30  | 87            | 95 (76)  | 61  |
| 200 | 46          | 87 (8)  | 43  | 97            | 100 (88) | 75  |
| 500 | 66          | 98 (3)  | 65  | 100           | 100 (97) | 97  |

Table 2: Validation sets of  $n$  items drawn in proportion to official weight (2,000 draws). “All”: share of selections equal to the full-benchmark winner. “Cert.”: the same among selections whose top-two gap on the validation set exceeds 2 points (share of draws in brackets). “Not”: among the rest.

while keeping at least 89.5% of the benchmark. On RewardBench 2 the median is 1.00 and the maximum 1.35. A curator who also keeps each section’s share, leaving no trace in the per-section breakdown, pays a median of 1.39 times the gap on RewardBench and 1.12 on RewardBench 2.

**Possible winners.** Because the bound is tight, the bracket on  $P(\delta)$  is narrow (Table 1). On RewardBench a pipeline that may have curated 1% of the benchmark can only say that one of 3 models is best, and at 5% one of 11. RewardBench 2’s winner is unique up to a 5% budget.

**Certificates on small validation sets.** A pipeline usually scores candidates on a small

validation set, where the certificate is the top-two gap on that set. Table 2 draws such sets at random, so that a set’s mean is an unbiased estimate of the official score. On RewardBench, a 100-item set selects the full-benchmark winner in 35% of draws; with 200 items, the 8% of selections certified at 2 points agree with it 87% of the time, against 43% for the rest. On RewardBench 2, whose winner leads by 5.22 points, most selections are certified and nearly all are correct. The certificate thus separates selections a pipeline can trust from those it should not, using only the validation scores.

## 5 A Selection Protocol

For an automated pipeline that selects an evaluator: (i) fix and hash the validation items before any candidate is scored, and log every filter applied to them; (ii) set a curation budget  $\delta$  for how much of the set could have been dropped or filtered; (iii) report the certificates  $g^*$  and  $k^*$  with the selection, and accept the winner only if  $g^* > \delta$ ; (iv) otherwise report the possible-winner set, bracketed as above, rather than a single winner. The certificates cost one pass over the score table, so the protocol adds no model calls.

## Limitations

We study one curation operation, removing items, on two reward-model benchmarks. Reweighting the kept items, paraphrasing them or changing the prompt format would give a curator more freedom, so the certificates hold only against removal. The item certificate is exact; the LP costs are achievable but may not be minimal. The validation-set experiment draws random sets, not adversar-

ial ones, and uses one threshold (2 points). We do not claim that any published result was curated. RewardBench 2’s Ties subset is excluded, so its scores differ slightly from the official ones.

## Reporting and Reproducibility

All results come from public per-item result files (allenai/reward-bench-results, allenai/reward-bench-2-results). All analysis ran on one laptop CPU with SciPy’s HiGHS solver; each full LP sweep takes one to two hours, the certificates seconds, and no models were run. One failed run is worth reporting: the first section-preserving search began its bisection at a very small kept share, where rounding can empty a section, and wrongly reported the leader as not crownable; the corrected search scans from  $f = 1$  downward. Code and derived data are at <https://github.com/adrian-erlikhman/curated-to-win>.

## Use of AI Assistants

An AI assistant (Claude, Anthropic) helped write the analysis code and draft the text. The author chose the study, checked the proofs and the results, and takes responsibility for the content.

## References

- Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*.
- Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, et al. 2021. The benchmark lottery. *arXiv preprint arXiv:2107.07002*.
- Ivan Dubrovsky, Anastasia Orlova, Illarion Iov, Nina Gubina, and Irena Gureeva. 2025. Selective adversarial attacks on LLM benchmarks. *arXiv preprint arXiv:2510.13570*.
- Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*.
- Polina Gordienko, Georg Schollmeyer, Frauke Kreuter, and Christoph Jansen. 2026. How hard is it to rig a benchmark? A social choice analysis of leaderboard robustness. *arXiv preprint arXiv:2605.23628*.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, et al. 2024. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, et al. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.
- Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, et al. 2025. RewardBench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*.
- Hosna Oyarhoseini, Jimmy Lin, and Amir-Hossein Karimi. 2026. A unified perturbation framework for analyzing leaderboard stability and manipulation. *arXiv preprint arXiv:2605.15761*.
- Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D’Souza, et al. 2025. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*.
- Guanhua Zhang and Moritz Hardt. 2024. Inherent trade-offs between diversity and stability in multi-task benchmarks. *arXiv preprint arXiv:2405.01719*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.