CORRELATION DOES NOT CERTIFY A POLICY CHOICE: WHAT BENCHMARKS SHOULD PUBLISH INSTEAD

Anonymous authors
Paper under double-blind review

ABSTRACT

A world model that correlates at .97 with real robot trials can still pick a worse policy per task than the best single policy on that board. Evaluators like it are validated against that printed table of real outcomes by publishing a correlation; we compute what one certifies about the choice. That certificate is an attained maximum over a cone of score matrices, depending on the decision.

For ranking the policies and shipping one, the usual claim, four of eleven stay within the cost of shipping the best one, and shortlisting two for hardware improves eight pairs. For choosing a policy per task, a decision we price against a correlation rather than observe, exactly one is adequate; ten leave open a regret above that baseline and two certify nothing at all.

A slightly higher correlation would not rescue the finer decision. Two numbers per task and two for the board bracket the correlation ruling out a five-point regret, wherever the ranges clear the budget. That bar lies between .978 and .998 here and above .97 on 498 of 499 random boards, and its mean did not fall as more tasks of the same kind were added.

A different summary converts better, and two of these benchmarks already publish it. An evaluator whose margin-weighted rank violation is v over K policies gives up at most Kv choosing a policy per task, and half that where its taskwise argmax is unique. Priced exactly, the statistic beats the correlation on all eleven pairs, but for a board-plus-noise evaluator it does nearly as well: the finding is the conversion, not the corpus. With no board, the correlation's bound gives nothing on six settings of thirteen, the statistic's on one. We recommend publishing it, with trial counts, per-task margins and ranges, and a declared accuracy.

1 INTRODUCTION

A world model that correlates at .97 with real robot trials can tell you which policy to ship: on three of the nine boards here, that correlation rules out any aggregate regret at all (App. A.2). It does not tell you, on any of those three, which policy to use on each task. Both halves of that claim are computable exactly, and this paper computes them. Using the world model per task implicitly claims to beat the board's best fixed policy, the single policy used on every task. At that correlation, even on the board it was validated against, the per-task choice can cost more. A rank-violation statistic that benchmarks already print does bound what that choice gives up, at most K times its value (Sec. 4).

The evidence in question is a correlation. SIMPLER reports how closely its simulated success rates track real robot trials (Li et al., 2024), WorldGym a correlation for a learned world model (Quevedo et al., 2026) and AutoEval for an autonomous real-robot proxy (Zhou et al., 2025). Ctrl-World and AgentRewardBench print the two columns and no correlation (Guo et al., 2026; Lù et al., 2025), and r is recomputed here for all eleven pairs. The practice recurs for judges of language models (Dorner et al., 2025; Kocmi et al., 2021) and for proxy metrics in online experimentation (Tripuraneni et al., 2024). The number then licenses a decision: a policy to ship, a checkpoint to keep, a candidate to run on hardware.

Fix a published outcome table p of K policies on T tasks (a board), and let an evaluator choose, on each task, the policy its scores rank first. The largest regret that choice can incur at correlation r with p is the exchange rate $\varepsilon^\star(r)$: a maximum over score matrices, attained, free of any distributional assumption.

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.

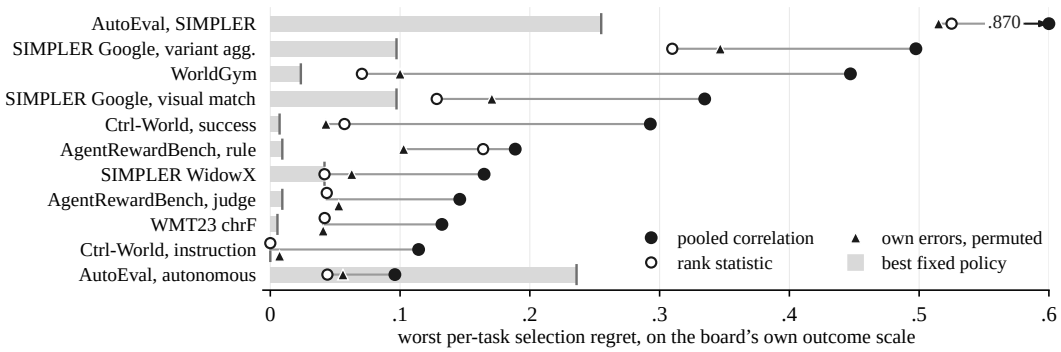


Figure 1: What the published evidence certifies about the decision. A filled circle is the worst the pooled correlation allows, an open circle the worst a rank-violation statistic allows, and a triangle the worst over the evaluator's own errors, permuted within each task. Marks that coincide are parted vertically. The band runs to the regret of shipping the board's best fixed policy, so a mark inside it means the evidence resolves the decision.

The exchange rate depends on the decision. These correlations can be adequate for ranking the policies and shipping one, the use Ctrl-World claims for its evaluator. Our corpus holds eleven evaluator-board pairs from four robotics benchmark suites, a web-agent benchmark and one machine-translation campaign. For shipping one, four stay within the best fixed policy's per-task regret, two of them at zero regret outright. Shortlisting two for hardware improves eight (Sec. 2). For choosing a policy per task the same correlations leave open a regret above that baseline on ten pairs (Fig. 1) and above random choice on nine, and on two they certify nothing at all. The evaluators themselves lose between zero and .415, and on four of the eleven more than the board's own best fixed policy. One is adequate, AutoEval's autonomous real-robot proxy, while the same benchmark's SIMPLER bridge on the same real cells certifies nothing. Restricted to the evaluator's own errors permuted within each task (the triangles), the worst case keeps .062 to .697 of its value. It still exceeds the baseline on the same ten pairs, nine if its members must stay in $[0,1]$ (Sec. 2, Prop. 11).

The bar is high. The correlation that rules out a regret above a tolerance ε (the bar r_ε) is bracketed by two numbers per task, the top-two margin g_t and the range L_t, with the policy count and one number for the board. A min-knapsack gives a correlation that fails, and a discrete program gives one above which every correlation succeeds; with three policies the first is the bar. At a five-point tolerance the bar is at least .978 on every board here. In our scan, drawing more uniform random tasks never lowered the mean bar, which converges (Sec. 3). Replicating a board cannot lower the bar at all (Prop. 7). The same margin profile bounds what an uncertain board can support (Lem. 2). Of the four designs stating a trial count, three, resampled at that count, would exceed a five-point regret more than once in a hundred (Sec. 4).

A published statistic that does convert. A margin-weighted rank-violation statistic converts to a per-task regret guarantee by a factor of K, and of $K/2$ when the published score table has a unique taskwise argmax. The factor is attained, so no constant improves it. That closed form needs no board and is weaker for it. SIMPLER reports one beside its correlations (Li et al., 2024), AutoEval follows it (Zhou et al., 2025), and Wang et al. (2026) adopt the practice, none of them converting it. The exact certificate the statistic licenses beats the correlation on all eleven pairs, by a median factor of 2.9 over the ten where it is positive, while the closed form loses on two of them. A plain count of the tasks on which the evaluator leads the board beats the correlation on ten (Props. 3 and 4). For evaluators drawn as the board plus noise the statistic wins nearly as often and by nearly as much, so the finding is the conversion, not the corpus. The gap is widest where no board is published. Benchmarks should publish the statistic, with trial counts, per-task margins and ranges, and a declared per-cell accuracy (Sec. 4).

2 What a published correlation certifies

Setting. A *board* is a table $p \in [0,1]^{K\times T}$ of real outcomes, K policies by T tasks, printed by the benchmark that owns it. Write $p_{(1)t} \geq \cdots \geq p_{(K)t}$ for the sorted column t, $g_t = p_{(1)t} - p_{(2)t}$ for its

Table 1: What a published correlation certifies, under the decisions a benchmark is used for. Each r is the Pearson correlation of the two printed columns, recomputed so the eleven are comparable, and $\varepsilon^\star(r)$ the worst regret an evaluator at that correlation can incur on the board. Under *per task* it picks a policy per task; under *aggregate*, the one it scores highest. *Shortlist* keeps the k it ranks highest. The table gives the shortlist's $\varepsilon^\star(r)$ at $k=2$, the smallest shortlist $k^\star$ meeting a five-point tolerance, and the bar $r_{k^\star}$ that shortlist sets on the board, .000 where no subset of that size is costly. *Realised* is what the published evaluator itself loses; *best fixed*, *random* and *worst possible* are the per-task regret of the board's best fixed policy, of choosing uniformly, and of the worst choice the board admits. The first is also the upside of the finer decision, and it is below a fortieth on six of the eleven pairs. $r_{.05}$ is that board's exact per-task bar, the correlation Sec. 3 brackets and Table 2 prices. The two bars are rounded up, so a printed requirement never understates the exact one; every other value is rounded to the places shown, halves upward.

		per task		aggregate		shortlist			per-task comparators			
evaluator and board	r	$\varepsilon^\star(r)$	realised	$\varepsilon^\star(r)$	realised	$k=2$	$k^\star$	$r_{k^\star}$	best fixed	random	worst possible	$r_{.05}$
WorldGym	.785	.447	.012	.000	.000	.000	1	.763	.024	.366	.632	.9888
SIMPLER Google, visual match	.941	.335	.017	.087	.000	.074	3	.914	.097	.291	.695	.9979
SIMPLER Google, variant agg.	.834	.498	.197	.391	.087	.087	4	.751	.097	.291	.695	.9979
SIMPLER WidowX	.841	.165	.042	.061	.000	.000	2	.750	.042	.150	.306	.9821
Ctrl-World, instruction	.968	.114	.000	.000	.000	.000	1	.905	.000	.250	.479	.9894
Ctrl-World, success	.832	.293	.007	.279	.000	.000	2	.753	.007	.238	.450	.9874
AutoEval, autonomous	.992	.096	.000	.112	.004	.004	2	.922	.236	.491	.744	.9979
AutoEval, SIMPLER	.450	.870	.415	.615	.000	.615	5	.000	.255	.583	.870	.9979
AgentRewardBench, judge	.965	.146	.000	.155	.000	.070	3	.963	.009	.097	.189	.9931
AgentRewardBench, rule	.931	.189	.071	.155	.125	.125	4	.000	.009	.097	.189	.9931
WMT23 chrF	.781	.132	.011	.119	.007	.049	2	.762	.006	.049	.136	.9790

top-two margin, $L_t = p_{(1)t} - p_{(K)t}$ for its *range*, and $S = \sum_{i,t}(p_{it} - \bar{p})^2$ about the mean entry $\bar{p}$. An *evaluator* is a second table y over the same cells, inducing $a(t) = \arg\max_i y_{it}$, whose regret is

$$R(a) \;=\; \frac{1}{T}\sum_{t=1}^{T}\big(p_{(1)t} - p_{a(t)t}\big). \tag{1}$$

Let $\mathscr{C}_a = \{y : y_{a(t)t} \geq y_{it}\ \forall i,t\}$ be the cone of score tables that produce a, and $d(a) = \min_{y\in\mathscr{C}_a}\|y-p\|_F^2$ its squared distance from the board. The cone is a product of per-task order cones, so $d(a)$ decomposes by task and is computed by pool-adjacent-violators on each column (Robertson et al., 1988).

Proposition 1. *For $K \geq 2$ and any choice a,* $\max\{\mathrm{corr}(y,p) : y \in \mathscr{C}_a \text{ nonconstant}\} \leq \sqrt{1 - d(a)/S}$, *the maximum being attained; and the inequality is an equality whenever $d(a) < S$.*

The maximum is over the closed cone, whose projection ties the selected policy with a leader on every conceded task, so an evaluator with strict taskwise maxima attains the value only in the limit there. Read backwards, Prop. 1 places the cone of an evaluator of correlation r within squared distance $S(1-r^2)$ of the board: no choice whose cone lies further away can produce that correlation. The *exchange rate* is the worst regret that survives:

$$\varepsilon^\star(r) \;=\; \max\big\{R(a) \;:\; d(a) \leq S(1-r^2)\big\}, \qquad r \in [0,1]. \tag{2}$$

For a summary s confining the evaluator to a set $\mathscr{S}(s)$, write $\varepsilon^\star(s) = \max\{R(a) : \mathscr{C}_a \cap \mathscr{S}(s) \neq \emptyset\}$; Sec. 4 prices two. Both $d(a)$ and $R(a)$ are additive over tasks, so (2) is a shortest-path problem over per-task rungs, one per policy a task could concede to. Solving it exactly in rational arithmetic returns the attaining score table with the value. Proposition 1 is Moreau's decomposition with Pythagoras, true of any closed convex cone holding the constants; the product structure over tasks and the boundary at $d(a) = S$ are ours. Apps. A and B prove every statement of this paper.

Two decisions. A benchmark's own claim is usually about ranking: Ctrl-World reports that its world model "can accurately rank policy performance" (Guo et al., 2026). Ranking and shipping one policy is the *aggregate* decision, whose cone constrains only the evaluator's policy means $\bar{y}_i$ and whose regret is the shortfall in mean outcome, $R_{\mathrm{agg}}(i) = \max_j \bar{p}_j - \bar{p}_i$. The two regrets differ by a constant: deploying policy i everywhere costs $R_{\mathrm{agg}}(i)$ plus the best fixed policy's own per-task regret. Judges, routers (Ong et al., 2025) and best-of-n selectors make the finer *per-task* choice that (1) scores, off-policy-evaluation benchmarks score it directly (Fu et al., 2021), and on WMT23 it is the choice of a system per language pair. RoboRouter routes per task on a robot board, beating the

best fixed policy on its own table of real hardware outcomes, 47 against 34 on five tasks at twenty episodes (Chen et al., 2026). We price that decision against a validation correlation rather than observe it. On the pairs that are not robotic it is the decision a practitioner makes, and on WMT23 and on AgentRewardBench's rule-based checker the published evaluator loses more than its board's best fixed policy. One correlation certifies the two decisions differently.

The audit. Table 1 applies (2) to eleven evaluator-board pairs from three domains. Eight evaluate robot policies: WorldGym's world model, Ctrl-World's under two criteria, SIMPLER's two protocols on a Google Robot and one on a WidowX arm, and AutoEval's autonomous real-robot proxy and SIMPLER bridge (Quevedo et al., 2026; Guo et al., 2026; Li et al., 2024; Zhou et al., 2025). Two are web agents, AgentRewardBench's four agents scored by a judge and by rule-based checkers (Lù et al., 2025); one is machine translation, WMT23's systems scored by human assessors and by chrF (Kocmi et al., 2023; WMT Conference, 2023). App. C gives the nine boards' provenance, reservations, shapes and counts.

For the aggregate decision, two of the eleven certify zero regret outright, and two others certify no more than the best fixed policy's own per-task regret. So on four of eleven, trusting the evaluator at worst doubles the best fixed policy's own regret; on the other seven it can more than double it.

For the per-task decision ten of the eleven are inadequate. Their $\varepsilon^\star(r)$ exceeds the best fixed policy's regret by a factor of at least 3.4, and the random policy's regret on nine. Two of the ten certify nothing at all: each admits the worst regret any choice on its board can incur. A correlation certifies nothing whenever its square is at most the share of S lying between the task means, a threshold the board fixes with no evaluator in hand (Cor. 4). WMT23's certificate reaches .97 of its board's worst possible regret. The gap to the evaluator's own regret is not slack. The maximum in (2) is attained, so each pair admits a score table, at or above its own recomputed correlation, whose per-task regret equals the certified value. The weakness is therefore in the correlation, not in these sources' evidence. The one adequate pair is AutoEval's autonomous proxy: its $r = .992$ is the highest in the corpus, and it certifies .096 against the best fixed policy's .236. Even so, its correlation falls short of the five-point bar of Sec. 3, which this board sets at .9979. Every certificate here conditions on the printed outcome table; none of them bounds regret against unknown true rates.

A decision these simulators can serve. A laboratory can instead shortlist, keeping the k policies the evaluator ranks highest overall, testing those on hardware and shipping the best. The tables producing a given shortlist form a closed convex cone, so Prop. 1 prices that decision unchanged. Its regret is the shortfall of the best shortlisted policy, and at $k = 1$ the decision is the aggregate one. The same eleven do far better here. The certificate cannot rise with k (Prop. 9), and eight pairs improve from $k = 1$ to $k = 2$ while three stay level. Keeping two certifies zero regret on four. Those four are exactly the corpus's four three-policy boards, where a shortlist of two keeps two thirds of the field. On the seven pairs whose boards have more it leaves .004 to .615 (Table 1). In practice a laboratory needs the smallest shortlist its correlation certifies at five points: one to five policies here, median two. Table 1 gives it with the bar that shortlist sets on the board, which a benchmark could publish. What these correlations license is screening rather than selection.

One board, two evaluators, two conventions. AutoEval's two bridges share four tasks with identical real cells. On those their per-task certificates differ by a factor of 14.5 with board, policies, tasks and decision held fixed, so that difference is the evaluator's alone (App. A.2). The second difference is in the statistic. Every r here is the Pearson correlation of the two printed columns pooled over all cells. SIMPLER computes one within each task and averages, and AutoEval follows it. The two confine the evaluator to different sets, so each has its own exchange rate, the second a knapsack over the same rungs (Prop. 10). Pricing each pair at the within-task convention instead moves the certificate by a factor between .250 and 2.334 and changes no pair's verdict against the best fixed policy. The within-task convention is the tighter summary on seven of eleven pairs; a pooled coefficient also credits ranking the tasks, which is not the decision. The .942 AutoEval states certifies .224, still below its best fixed policy's .236.

What the evaluators' own accuracy leaves. A maximum over every real matrix invites the objection that its maximiser is not a plausible evaluator, and the sources print the evaluator column, so it is testable. Write δ for the largest per-cell error after the best affine calibration. On the corpus δ

runs from .091 on WMT23 to .472 on AutoEval's SIMPLER bridge. Capping at each pair's own δ still leaves the worst case equal to $\varepsilon^\star(r)$ on four pairs and at least .586 of it on the rest (Prop. 5).

Restricting to the evaluator's own errors. Shrink that maximum to a class the published evaluator generates. Fix each task's errors $e = y - p$ and let π_t permute which policy carries which error on task t, whose columns are $p_{\cdot t}$ and $e_{\cdot t}$. Every member reproduces that task's error magnitudes exactly and differs from the published evaluator in attribution alone. The class is finite, indexed by $(K!)^T$ profiles, and contains the evaluator itself. A within-task permutation preserves every column sum and every column's residual multiset, so the pooled correlation depends on the permutations through the single scalar $C = \sum_t \langle p_{\cdot t}, \pi_t(e_{\cdot t})\rangle$, which is additive over tasks as the regret is. The set failing r is then one interval in C and the maximisation an integer knapsack, solved exactly in rational arithmetic (Prop. 11). The value is taken over the class as written; kept inside $[0,1]$, SIMPLER WidowX's value falls to its best fixed policy's regret (App. A.6). On the eleven pairs it keeps between .062 and .697 of $\varepsilon^\star(r)$, median .381 (the triangles of Fig. 1). The conclusion survives: the restricted value stays above five points on eight pairs and exceeds the best fixed policy's regret on ten of the eleven, the same ten, and on eight with one permutation for every task, five if its members must stay in $[0,1]$. The restriction a reader may try first, a board plus Gaussian noise, is no restriction at all: its support contains the unrestricted maximiser on every pair here, so it returns $\varepsilon^\star(r)$ unchanged.

Figure 2 in App. A traces $\varepsilon^\star(r)$ for the per-task decision, on an axis logarithmic in $1-r$. Each curve drops to its own best-fixed-policy line only past a crossing. The crossings, each rounded up to four places, run from .9521 to .9997 over the nine boards, and exactly one correlation lies above its own.

3 The bar an evaluator must clear, from the margins and the ranges

If Table 1's certificates for the per-task decision are weak, how strong a correlation would have made them strong? Fix a tolerance ε on that decision's regret and let

$$d_\varepsilon = \min\{d(a) : R(a) > \varepsilon\}, \qquad r_\varepsilon = \sqrt{1 - d_\varepsilon/S}, \tag{3}$$

so that every choice an evaluator with $\mathrm{corr}(y,p) > r_\varepsilon$ can produce has regret at most ε. We call r_ε the *bar*. From four policies on, computing it exactly needs the whole board, though the bracket below closes on it on six of the nine boards here. The two results below bracket it from the per-task margins and ranges, the policy count and one number for the whole board, which is all a benchmark would have to print. Everything below rests on one calculation: write $d_t(i)$ for the squared distance from column t to the set where policy i is a maximum, and $\ell = p_{(1)t} - p_{it}$ for its loss.

Lemma 1 (per-task rung cost)**.** *$d_t(i) \geq h_t(\ell)$ for every policy and every column, where $\ell = p_{(1)t} - p_{it}$, $h_t(\ell) = \ell^2/2$ for $\ell \leq 2g_t$ and $h_t(\ell) = \ell^2/2 + (\ell - 2g_t)^2/6$ above. It is convex, increasing, never below $\ell^2/2$, and reads the column only through g_t. When $\ell > 0$ equality holds exactly when every entry other than p_{it}, one leader and one runner-up is at or below the pooled level of App. A.1. If i is a runner-up it is unconditional: $d_t(i) = g_t^2/2$.*

The projection pools i with the entries above that level: the leader always, the runner-up only once ℓ passes $2g_t$. Below that, the midpoint of i and the leader dominates it. So for a runner-up nothing else moves, and regret is cheapest, per unit, on the smallest-margin tasks.

Theorem 1 (necessary side)**.** *Suppose $\sum_t L_t > T\varepsilon$, and let u_t be the cone cost of the bottom entry of column t once its $K-2$ middle entries are raised to the runner-up: $u_t = L_t^2/2$ when $L_t \leq 2g_t$ and $u_t = (K-2)g_t^2 + L_t^2 - ((K-2)g_t + L_t)^2/K$ above. Then $d_\varepsilon \leq D^+(\varepsilon)$, where*

$$D^+(\varepsilon) = \min\Big\{ \tfrac{1}{2}\textstyle\sum_{t\in U} g_t^2 + \sum_{t\in B} u_t \;:\; U \cap B = \emptyset,\ \sum_{t\in U} g_t + \sum_{t\in B} L_t > T\varepsilon \Big\}, \tag{4}$$

and consequently $r_\varepsilon \geq \sqrt{\max\{0, 1 - D^+(\varepsilon)/S\}}$. Here $d_\varepsilon = D^+(\varepsilon)$ exactly when some minimiser of (3) *demotes each task it concedes to a runner-up or to a bottom policy of cone cost u_t, which every bottom policy has when $L_t \leq 2g_t$; when $K \leq 3$ it is automatic.*

Raising an entry the choice does not take can only raise a cone cost, and every middle entry sits at or below the runner-up. So u_t bounds the cost of an actual choice, the demotion to a bottom policy.

With three policies every policy is a leader, a runner-up or a bottom one and u_t is exact, so (4) is d_ε itself. From four policies on it can be loose by an unbounded factor, yet it is the smallest upper bound on d_ε the margins, the ranges and K allow, and, save at a boundary, Thm. 2's program is the largest lower one (App. A.1).

Theorem 2 (sufficient side)**.** *Suppose $\sum_t L_t > T\varepsilon$. Then $d_\varepsilon \geq D^{\mathrm{disc}}(\varepsilon)$, where*

$$D^{\mathrm{disc}}(\varepsilon) = \min\Big\{\textstyle\sum_t h_t(\ell_t) \;:\; \ell_t \in \{0\} \cup [g_t, L_t],\ \sum_t \ell_t \geq T\varepsilon\Big\}, \tag{5}$$

and consequently $r_\varepsilon \leq \sqrt{1 - D^{\mathrm{disc}}(\varepsilon)/S}$. Writing $f_t(\ell) = g_t \ell$ for $0 \leq \ell \leq g_t$ and $f_t(\ell) = \ell^2$ above, the convex envelope program $D^-(\varepsilon) = \frac{1}{2}\min\{\sum_t f_t(\ell_t) : 0 \leq \ell_t \leq L_t,\ \sum_t \ell_t \geq T\varepsilon\}$ satisfies $D^-(\varepsilon) \leq D^{\mathrm{disc}}(\varepsilon)$, so it is a weaker endpoint of the same kind; its minimiser is a water-filling level at a common marginal price, and App. A.1 gives the level per task and the price.

By Lemma 1, conceding loss ℓ on task t costs at least $h_t(\ell)$, and no choice can concede a loss strictly between 0 and g_t, since the second-best outcome is the first one below the leader. So the losses of any costly choice are feasible for (5), and the endpoint keeps each task's margin and range and forgets the rest of its column. It is also not convex, its feasible set omitting $(0, g_t)$ on each task, and setting $g_t = L_t$, as every two-policy board does, turns it into a min-knapsack over the same squared margins. It is solved not by water filling but by a search over which tasks are conceded. Each conceded set leaves a strictly convex separable objective under per-task bounds and one linear constraint, whose minimiser prices the free tasks at one common marginal charge. Two facts prune the search. A conceded task costs at least half its squared margin, which is what h_t charges at a runner-up demotion, so the running total only grows. And $D^+(\varepsilon)$ supplies a starting value, its minimiser being feasible here at no greater cost. From four policies on, Table 2 uses (5) itself. On several hundred random boards, each priced at several tolerances, only the envelope program stays affordable, so the random scans below use it from four policies on. $D^-(\varepsilon)$ can therefore fall below the discrete minimum it relaxes, and the endpoint is conservative by that much.

Corollary 1. *Suppose $\sum_t L_t > T\varepsilon$. Then $\sqrt{\max\{0, 1 - D^+(\varepsilon)/S\}} \leq r_\varepsilon \leq \sqrt{1 - D^{\mathrm{disc}}(\varepsilon)/S}$, and both endpoints are functions of $(g_t)_t$, $(L_t)_t$, K and S alone. Neither program reads S: it enters once, in the square root. When $K \leq 3$ the left inequality is an equality, so the left endpoint is sufficient as well and the bracket closes on r_ε; where $\sum_t L_t \leq T\varepsilon$ no correlation is required at all.*

The bracket is two-sided for a reason: its lower endpoint is necessary and its upper endpoint sufficient, and only the upper endpoint licenses calling an evaluator adequate. A correlation below the lower endpoint demonstrably fails, (4) exhibiting the costly choice; one above the upper endpoint demonstrably succeeds. A board could withhold everything but $2T + 1$ numbers and its policy count and still be bracketed: (4) and (5) read only the margins, the ranges and K, so two boards that share them pose the same two programs. From four policies on the exact bar has no such property, as the released code checks on pairs of boards built to share them and S.

The bar is high, and more tasks of one kind did not lower its mean. Table 2 in App. A evaluates the bracket at $\varepsilon = .05$ on the nine distinct boards. It contains the exact bar on every one, at an exact width running from zero to .0257. Thm. 2 gives $D^{\mathrm{disc}}(\varepsilon) \leq d_\varepsilon$ and Thm. 1 gives $d_\varepsilon \leq D^+(\varepsilon)$. Two endpoints that meet therefore pin the exact bar, from $(g_t)_t$, $(L_t)_t$, K and S with no outcome table read. The widest bracket is WMT23's, $[.9553, .9810]$ around an exact .9790, the lowest bar in the corpus: its margins total .079 against a budget of .400.

On 499 random boards of independent uniform cells with $K \in \{3, \ldots, 6\}$ and $T \in \{4, \ldots, 20\}$, one of 500 draws being discarded (App. A.3), the bracket contains the bar every time, at a median width of .0008. The bar exceeds .97 on 498 of them. Under three structured laws the bar clears .97 on at least .970 of 20000 draws under each, the interval taken simultaneously over the three (App. A.3). Here a five-point decision needs a correlation near .99, and .97 is not close.

Theorem 1's runner-up part explains this, and why more tasks of one kind need not help. Setting $B = \emptyset$ in (4) leaves the runner-up program $D_g^+(\varepsilon) = \frac{1}{2}\min\{\sum_{t \in U} g_t^2 : \sum_{t \in U} g_t > T\varepsilon\}$, a weaker necessary endpoint that exists where $\sum_t g_t > T\varepsilon$. Let $U^\star$ be an inclusion-minimal minimiser, the one the exact program returns, let $m^\star = \sum_{t \in U^\star} g_t$ be its margin mass, $\mu(\varepsilon) = \big(\sum_{t \in U^\star} g_t^2\big)/m^\star$ and $\sigma^2 = S/(KT)$:

Proposition 2. *Assume $\sum_t g_t > T\varepsilon$. If $U^\star$ is inclusion-minimal then $T\varepsilon < m^\star \le T\varepsilon + \min_{t\in U^\star} g_t$, and therefore*

$$\frac{\varepsilon\,\mu(\varepsilon)}{2K\sigma^2} \;<\; 1-\Big(\sqrt{1-D_g^+(\varepsilon)/S}\Big)^2 \;\le\; \Big(\varepsilon+\tfrac{1}{T}\min_{t\in U^\star} g_t\Big)\frac{\mu(\varepsilon)}{2K\sigma^2}. \tag{6}$$

The task count appears only in the relative width, which is $O(1/(T\varepsilon))$. It cancels from the leading term because the regret target is a per-task mean while S grows with T, so d_ε and S both grow in proportion to T. Draw the columns independently from a fixed law whose top-two margin has a density, and fix $0<\varepsilon<\mathbb{E}[g]$: then $\mu(\varepsilon)$, σ^2, D_g^+/S and D^-/S converge. The bar converges too, and on the evidence of App. A.3 to neither of their limits.

Fig. 3 in App. A shows the convergence. As the same board grows, the mean bar rises at every step and the steps shrink. The smallest step clears 4.3 standard errors of its own paired difference. Across that one scan's 499 random boards the correlation between the bar and T is .32, and $-.07$ once the closed form is subtracted. The policy count is the dependence that matters. The coarser decision has a bar too, exact rather than bracketed, and read off the policy means with the board's shape and S (Cor. 3). It runs from .7629 to .9975 across the corpus, lying below the per-task bar on eight pairs and above it on three, and two of the eleven correlations clear their own where none clears the per-task bar.

The tolerance depends on the outcome scale. Five points of mean task success on a robot board and five points of a hundred-point human judgement are not the same demand. On WMT23 that budget is five times the board's total margin mass, the source of its low bar. Every board should be read against a tolerance its own scale licenses, as App. A.3 reads this one. We fix one so the columns are comparable.

When the board itself is uncertain. The printed board, the anchor every certificate here is read against, is an estimate from finitely many trials. Ask for a choice whose regret is at most ε for *every* board in the box of radius η around it, cell by cell. Everything then turns on one quantity, the hinge: the least regret that box forces on any choice. Lemma 2 in App. B computes it exactly. It never depends on the evaluator. Unless a cell's interval meets 0 or 1, it depends on the board only through its margins: a task contributes once 2η exceeds its margin. App. B tabulates the radii the corpus admits at a five-point tolerance. The hinge treats the declared radius as given; it does not establish that any published anchor is accurate to that radius. Section 4 supplies the only radius these sources license, their own sampling error, and composes the two.

4 What to publish

Sections 2 and 3 show that the correlation certifies little for the per-task decision and that none here clears the bar. That is a claim about the summary, not the board, where a full score table fixes the regret exactly. $\varepsilon^\star(s)$ answers which cheap summary carries most of it. All eight source tables print the full board, so for them this asks only for a compression. What follows binds in two cases: the three source tables that state no trial count, and the benchmarks that print a correlation and a statistic and no table.

A rank statistic converts linearly. SIMPLER publishes one, saying a correlation "only assesses the linear fit between real and simulated performance" (Li et al., 2024). Its mean maximum rank violation charges each policy the largest real gap to any policy it is ranked against wrongly. That charge is averaged over policies and tasks:

$$\mathrm{MMRV}(y,p) = \frac{1}{KT}\sum_{t=1}^{T}\sum_{i=1}^{K}\max_{j:\,[y_{it}>y_{jt}]\neq[p_{it}>p_{jt}]}|p_{it}-p_{jt}|, \tag{7}$$

with $[\cdot]$ the indicator and the convention that a maximum over no violated pair is zero. A tie on one side of a strict board ordering is charged to the policy the board ranks higher, and to it alone.

Proposition 3. *Let y satisfy $\mathrm{MMRV}(y,p)\le v$ and let a take a taskwise maximiser of y. Then* (i) *$R(a)\le Kv$, unconditionally; and* (ii) *if y scores one policy strictly first on every task, $R(a)\le \frac{K}{2}v$. In each case the worst regret compatible with a published v is an exact knapsack over the per-task charges, and attaining the factor requires every conceded task to be conceded to its runner-up.*

Every policy the board ranks above the selected one is charged its own gap, whether separated or tied, so the task costs at least the loss. If the scores put the selected policy *strictly* first, that policy is charged the largest of those gaps too, and the task costs at least twice the loss. An evaluator that ties escapes the second charge, and the factor doubles.

Table 3 in App. A prices the three summaries. Nine of the eleven published score tables have a unique taskwise argmax, so the tighter knapsack applies; the other two get the unconditional one, which by itself beats the correlation on all eleven. Under the certificate each table licenses, the rank statistic beats the correlation by factors of 1.15 to 6.34, median 2.88 over the ten pairs with a positive certificate, and without bound on the eleventh. Those are exact knapsacks; the closed form is weaker, losing to the correlation on two pairs. On the nine pairs with a unique taskwise argmax, publishing the correlation beside the statistic adds nothing: the certificate under both at once is the statistic's own. Beating the correlation is not the same as resolving the decision: the rank statistic's certificate falls at or below the best fixed policy's regret on three pairs of eleven, against the correlation's one. Under the board-plus-noise null below, it does so on three or more pairs at least .379 of the time.

A count also beats the correlation. A benchmark unwilling to adopt a new statistic can report the number m of tasks on which its evaluator's choice leads the board.

Proposition 4. *With the ranges ordered $L_{[1]} \geq \cdots \geq L_{[T]}$, any choice that leads the board on at least m of T tasks has $R(a) \leq \frac{1}{T}\sum_{j \leq T-m} L_{[j]}$, and the bound is attained.*

The $T-m$ tasks that may go wrong lose at most their full ranges. Proposition 4 needs the published range profile and one integer, no score and no distributional assumption. It beats the correlation on ten of eleven pairs, an advantage the null below reproduces (App. A.4). The exception is SIMPLER's variant aggregation: despite a correlation of .834 it leads the board on only one task of four, so the count certifies .593 against the correlation's .498.

How much of that is due to these benchmarks. Little that we can show. Replace each pair's evaluator by the board plus Gaussian noise clipped to $[0,1]$, at the scale whose median correlation is that pair's recomputed one. The null's median factor is lower than the corpus's (App. A.4). Only WMT23's factor lies significantly above its null: 36 of 20000 null draws reach it (one-sided $p = .002$, .02 after Bonferroni over the ten pairs with a factor). Beyond that one pair, and with the rule-based checker's null draws priced as the checker is, neither comparison is evidence about these benchmarks. Drawing the boards at random too, over six noise scales that nearly span the corpus's own correlations, leaves the rank statistic the tighter of the two on at least .938 of draws at every scale, the interval taken simultaneously over the six. What separates the summaries is the conversion, which is a theorem.

What converts when no board is published. A benchmark that prints a correlation and a statistic and no table leaves two numbers and a policy count; both convert, unequally. Proposition 3(i) needs K and the statistic alone. The exact exchange rate needs the board; without one, the certificate a correlation carries is at most the crude bound of Cor. 2, vacuous at or below $r = \sqrt{1-2/K}$. Table 7 prices the thirteen settings Wang et al. (2026) publish that way, five policies in each. That bound is vacuous on six of the thirteen; the rank statistic's is vacuous on one and the tighter on twelve. On the setting whose published statistic is zero it certifies zero regret outright. The published r and MMRV order the settings alike, at $-.94$ on opposite scales; they differ in sharpness. That sharpness is not an artefact of a loose bound. Wherever Cor. 2 is not vacuous at $K = 5$, App. A.4 exhibits boards attaining at least $\sqrt{3/5}$ of it. No sharper bound could therefore close a gap of a factor of five or more.

The boards mostly do not resolve their own margins. Five of the eight source tables state a per-cell count, covering four independent designs (App. B), from ten trials on WorldGym to seventy-five on a SIMPLER task. Resample each board at its own stated depth, taking the printed rates as truth and every trial as independent, the most favourable dependence the designs permit, and choose per task from the resample. That choice's regret is an average of independent per-task losses, so the probability that it stays within five points is exact: .997 on AutoEval, .969 on WorldGym, .928 on SIMPLER WidowX and .909 on SIMPLER Google Robot. Only one of the four, AutoEval, whose every cell carries fifty trials, reaches .99. This refutes the retrospectively registered criterion "No board may reach .99 probability of selecting within the tolerance under a fresh read at the depth its own source states" (App. D). Reaching .99 would take between .76 and 2.67 times the published

depths: thirteen trials per cell on WorldGym, a uniform seventy-two on Google Robot. Of the thirty tasks on these four boards, thirteen have a top-two margin exceeding twice its own sampling standard error at its published depth, and on both SIMPLER boards none does.

Lem. 2 answers a related question at a stated confidence. Take each cell's exact binomial interval at its stated depth as a box of boards, not a plug-in standard error, which would assign radius zero to these four boards' thirty-two cells reporting no successes or no failures. The hinge is the least regret that box forces on any choice, so some choice is certifiable at five points exactly when the hinge is no larger. The interval's levels are per-cell marginal ones, so a simultaneous region would be wider. That could remove a certifiable choice but never add one. At the 68% level the hinge is .0651 on SIMPLER WidowX, .0437 on WorldGym, .0366 on SIMPLER Google Robot and .0236 on AutoEval. With counts rounded half to even, WidowX admits no certifiable choice at five points and the other three do. Only Google Robot and AutoEval do so under every count their rates admit: WorldGym's hinge runs from .0382 to .0653 over its half-trial credits, and WidowX's from .0486 to .0652 over its two irreconcilable rates (App. B). At 95% the four are .2166, .2090, .1445 and .0798. At that level none of the four admits a certifiable choice, for any evaluator, a perfect one included. Where that holds, the binding constraint is the trials, not the evaluator or any summary of it.

Recommendation. A benchmark publishing a validation correlation should publish, for that board:

1. The per-cell trial count behind each entry, which three of the eight source tables here do not state and without which no margin can be called resolved.
2. A margin-weighted rank-violation statistic, with one bit for whether its taskwise argmax is unique, or failing that the agreement count m (Props. 3 and 4). Any correlation reported beside them should say which one it is.
3. The per-policy mean outcome, the per-task top-two margin g_t and range L_t, and S. With the board's shape, the means and S give exactly the bar for the decision a benchmark usually claims (Cor. 3). With the same S, the margins and ranges bracket the bar for the finer one (Cor. 1), and the margins bound what an uncertain board can support (Lem. 2). Each is read at a tolerance the board's outcome scale licenses.
4. The per-cell accuracy δ claimed for the evaluator, which converts the correlation into a certificate (Prop. 5). A five-point certificate needs one 1.2 to 6.7 times finer than these evaluators attain.

The first is known to whoever ran the trials, and the other three total $K + 2T + 3$ numbers and one bit. The released library converts each, computing the bracket and the aggregate bar from those numbers alone. Together the four turn a validation table into a certificate, and with no board at all they are tighter than a correlation with the board on ten of the eleven pairs here (App. A.2).

5 Related work

Turning a validation statistic into a decision guarantee is old (Taylor & Russell, 1939; Cronbach & Gleser, 1965; Tripuraneni et al., 2024), and the surrogate-endpoint and proxy-reward literatures record the same lesson (Baker & Kramer, 2003; Laidlaw et al., 2025; Liu et al., 2026). All of it is in population, or, in Skalse et al. (2024), two-sided but tight only as a class. Ours is a single attained maximum over a printed table, at the statistic that benchmarks publish. Closest in purpose is Dorner et al. (2025), whose margin for trusting a judge's ranking does not soften as evaluation data grow; ours concerns per-task selection and is attained, not bounded. Badithela et al. (2026) instead bound the true rates from paired trials, a question this paper leaves aside. Appendix E gives the rest.

6 Scope and limitations

Each quantity here is a statement about a printed table. It does not bound regret on a new policy, a new task, or against the unknown rates the board estimates. The certificate is attained over every real score matrix at the recomputed correlation, a class far larger than the plausible simulators; one for per-cell unbiased errors needs a model, and we do not compute one. The per-task decision is priced against a correlation rather than observed. With three policies the bracket of Cor. 1 always closes on the bar; from four on, the necessary endpoint can be loose by an unbounded factor. The corpus is small and mostly robotic: nine boards, on five of which one policy leads at least three tasks in four, and two others share cells. The scan draws cells independently, with no policy or task effect.

REPRODUCIBILITY STATEMENT

The archive contains the manuscript, the eight source tables and one board-free transcription, the analysis code and the computed results. `MANIFEST.md` maps each numbered statement, table and figure to the script that computes it and the output it writes. Every analysis runs from `code/` without a robot, a GPU or network access. Sampling models are named beside the analyses that use them. No seed is printed in the paper; each is fixed in the generator script that uses it.

`regenerate.py` deletes the generated outputs in a scratch copy, reruns their generators, compares the numerical artifacts field by field and every figure pixel by pixel. `numeral_coverage.py` binds each central quantity in the manuscript to an exact field of a released artifact, and `SCOPING.md` binds the principal assumptions and limitations to the sentences that state them. App. C states which tables enter the corpus and which do not, and App. D inventories the released code and outputs. Reproducing these calculations does not establish that they hold outside their assumptions.

ETHICS STATEMENT

This work re-analyses published validation evidence. It collects no new data from humans or robots and involves no human subjects. Its findings concern what published numbers can and cannot establish, not the conduct of the authors who published them. The instruments it releases are meant to let a laboratory price an evaluator under stated assumptions before trusting it with a decision.

AI USE STATEMENT

Of the tasks with required disclosure, we used large language model assistants, under the authors' direction, to propose or refine hypotheses and to implement the methods in the released code. We did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide critical ingredients for proving those claims or to write the proofs, to design or give feedback on the research methodology or the experiments, to clean or reformat datasets, or to interpret results. Translation and qualitative or thematic data analysis are not applicable to this work.

Hypotheses. Assistants helped propose or refine hypotheses. The claims the paper states, and their proofs, are the authors'.

Code. Assistants helped implement the methods in the released code. That code is checked by the test suite and the regeneration checks the archive releases (App. D).

Writing. Assistants edited the text for clarity, grammar and concision, kept terminology and notation consistent across the body and the appendices, and checked that the prose, the captions and the tables describe each quantity in the same terms.

Drafting. Assistants drafted text for sections of the paper, including captions and the appendices that describe the corpus and the released code. The drafts were revised under the authors' direction, and the authors decided what the paper claims and how it is organised.

Verification. Assistants were used to check the work: re-deriving statements independently, testing them by exact computation on small cases, recomputing the reported numbers from the tables, and comparing citations with the works they cite. No data in the paper was generated by a language model. The corpus consists of published tables, and App. C states how their transcription is checked; every simulated quantity comes from the released code, with each seed fixed in the script that uses it.

The authors take full responsibility for every claim, derivation, number, reference and line of code in this work, including whatever an assistant helped produce.

REFERENCES

Hassene Aissi, Cristina Bazgan, and Daniel Vanderpooten. Min-max and min-max regret versions of combinatorial optimization problems: A survey. *European Journal of Operational Research*, 197 (2):427–438, 2009. doi: 10.1016/j.ejor.2008.09.012.

Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Cyrus Neary, Edward S. Hu, Kanav Arora, Kirsty Ellis, Luca Macesanu, Matthew Leonard, Meedeum Cho, Ozgur Aslan, Shivin Dass, Tony Wang, Xingfang Yuan, Abhishek Gupta, Dinesh Jayaraman, Glen Berseth, Kostas Daniilidis, Roberto Martín-Martín, Youngwoon Lee, Percy Liang, Chelsea Finn, and Sergey Levine. RoboArena: Distributed real-world evaluation of generalist robot policies. In *Conference on Robot Learning (CoRL)*, volume 305, pp. 336–364. PMLR, 2025. arXiv:2506.18123.

Apurva Badithela, David Snyder, Lihan Zha, Joseph Mikhail, Matthew O'Kelly, Anushri Dixit, and Anirudha Majumdar. Reliable and scalable robot policy evaluation with imperfect simulators. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. arXiv:2510.04354.

Stuart G. Baker and Barnett S. Kramer. A perfect correlate does not a surrogate make. *BMC Medical Research Methodology*, 3:16, 2003. doi: 10.1186/1471-2288-3-16.

Pierre Boyeau, Anastasios N. Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 5276–5290. PMLR, 2025. arXiv:2403.07008.

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910. doi: 10.1111/j.2044-8295.1910.tb00207.x.

Tomasz Burzykowski and Marc Buyse. Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*, 5(3):173–186, 2006. doi: 10.1002/pst.207.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020. doi: 10.18653/v1/2020.emnlp-main.745. arXiv:2010.06595.

André Chassein and Marc Goerigk. Minmax regret combinatorial optimization problems with ellipsoidal uncertainty sets. *European Journal of Operational Research*, 258(1):58–69, 2017. doi: 10.1016/j.ejor.2016.10.055.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. Prediction-powered ranking of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2402.17826.

Yiteng Chen, Zhe Cao, Hongjia Ren, Chenjie Yang, Wenbo Li, Shiyi Wang, Yemin Wang, Li Zhang, Yanming Shao, Zhenjun Zhao, Huiping Zhuang, and Qingyao Wu. RoboRouter: Training-free policy routing for robotic manipulation. *arXiv:2603.07892*, 2026.

Lee J. Cronbach and Goldine C. Gleser. *Psychological Tests and Personnel Decisions*. University of Illinois Press, Urbana, IL, 2nd edition, 1965.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.13341.

Othman El Balghiti, Adam N. Elmachtoub, Paul Grigas, and Ambuj Tewari. Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research*, 48(4):2043–2065, 2023. doi: 10.1287/moor.2022.1330. Preliminary version in NeurIPS 2019. arXiv:1905.11488.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1):9–26, 2022. doi: 10.1287/mnsc.2020.3922. arXiv:1710.08005.

Paul Embrechts, Giovanni Puccetti, and Ludger Rüschendorf. Model uncertainty and VaR aggregation. *Journal of Banking & Finance*, 37(8):2750–2764, 2013.

Thomas R. Fleming and David L. DeMets. Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine*, 125(7):605–613, 1996. doi: 10.7326/0003-4819-125-7-199610010-00011.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Tom Le Paine. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2103.16596.

Gemini Robotics Team. Evaluating Gemini Robotics policies in a Veo world simulator. *arXiv:2512.10675*, 2025.

Polina Gordienko, Georg Schollmeyer, Frauke Kreuter, and Christoph Jansen. How hard is it to rig a benchmark? A social choice analysis of leaderboard robustness. *arXiv:2605.23628*, 2026.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.10125.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022. doi: 10.1287/mnsc.2022.4383. arXiv:2011.03030.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.09144.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, volume 270, pp. 2679–2713. PMLR, 2024. arXiv:2406.09246.

Leslie Kish. *Survey Sampling*. Wiley, New York, 1965.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pp. 478–494. Association for Computational Linguistics, 2021. arXiv:2107.10821.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1/.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1999–2014. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.110. arXiv:2401.06760.

Anany Kotawala. Resolution diagnostics for paired LLM evaluation. *arXiv:2605.30315*, 2026.

Panos Kouvelis and Gang Yu. *Robust Discrete Optimization and Its Applications*. Kluwer Academic Publishers, 1997. doi: 10.1007/978-1-4757-2620-6.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.03185.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control (L4DC)*, volume 120, pp. 761–770. PMLR, 2020. arXiv:2002.04523.

Eddie Landesberg. When LLM judge scores look good but best-of-N decisions fail. *arXiv:2603.12520*, 2026.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning (CoRL)*, volume 270, pp. 3705–3728. PMLR, 2024. arXiv:2405.05941.

Mark Lindeman and Philip B. Stark. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy*, 10(5):42–49, 2012. doi: 10.1109/MSP.2012.56.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. Robust optimization for mitigating reward hacking with correlated proxies. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2604.12086.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. AgentRewardBench: Evaluating automatic evaluations of web agent trajectories. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.08942.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 34303–34326. PMLR, 2024. arXiv:2402.14992.

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

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv:2411.00640*, 2024.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016. arXiv:1511.04599.

Jean-Jacques Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 255: 238–240, 1962.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.18665, titled "RouteLLM: Learning to Route LLMs with Preference Data".

Ross L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989. doi: 10.1002/sim.4780080407.

Giovanni Puccetti and Ludger Rüschendorf. Computation of sharp bounds on the distribution of a function of dependent risks. *Journal of Computational and Applied Mathematics*, 236(7): 1833–1840, 2012.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.00613.

Tim Robertson, F. T. Wright, and Richard L. Dykstra. *Order Restricted Statistical Inference*. Wiley, New York, 1988.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. STARC: A general framework for quantifying differences between reward functions. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.15257.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3): 271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

H. C. Taylor and J. T. Russell. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 23(5):565–578, 1939. doi: 10.1037/h0057079.

Nilesh Tripuraneni, Lee Richardson, Alexander D'Amour, Jacopo Soriano, and Steve Yadlowsky. Choosing a proxy metric from past experiments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 5803–5812, 2024. doi: 10.1145/3637528.3671543. arXiv:2309.07893.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004. doi: 10.1214/aos/1079120131.

Joseph A. Vincent, Haruki Nishimura, Masha Itkina, Paarth Shah, Mac Schwager, and Thomas Kollar. How generalizable is my behavior cloning policy? A statistical approach to trustworthy performance evaluation. *IEEE Robotics and Automation Letters*, 9(10):8619–8626, 2024. doi: 10.1109/LRA.2024.3445635. arXiv:2405.05439.

Shuo Wang, Hanyuan Xu, Yingdong Hu, Fanqi Lin, and Yang Gao. A practical recipe towards improving sim-and-real correlation for VLA evaluation. *arXiv:2606.10366*, 2026.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05584.

WMT Conference. WMT23 general MT task: Primary system submissions, human evaluations and automatic scores. `https://github.com/wmt-conference/wmt23-news-systems`, 2023. File `scores/automatic-scores.tsv`; accessed 2026-09-17.

Lirong Xia. Computing the margin of victory for various voting rules. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*, pp. 982–999, 2012. doi: 10.1145/2229012.2229086.

Zeyu Xu, Xiaojie Mao, Hao Mei, and Yue Liu. Evaluating surrogates in individualized treatment rules. *arXiv:2512.00405*, 2025.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? A decision-making-centric position. *arXiv:2606.15032*, 2026.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-World: World models in a closed-loop world. In *International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2510.18135.

Kaifeng Zhang, Shuo Sha, Hanxiao Jiang, Matthew Loper, Hyunjong Song, Guangyan Cai, Zhuo Xu, Xiaochen Hu, Changxi Zheng, and Yunzhu Li. Real-to-sim robot policy evaluation with Gaussian splatting simulation of soft-body interactions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026b. arXiv:2511.04665.

Zhiyuan Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, and Sergey Levine. AutoEval: Autonomous evaluation of generalist robot manipulation policies in the real world. In *Conference on Robot Learning (CoRL)*, volume 305, pp. 1997–2017. PMLR, 2025. arXiv:2503.24278.

A Proofs and supporting tables

Standing conventions. App. A.1 proves the eight statements of the body in the order the body states them, and Apps. A.2 to A.4 support Secs. 2 to 4 in turn. App. A.5 prices coarser decisions and both conventions for a correlation, and App. A.6 prices the worst case over the evaluator's own errors. $D^{\mathrm{disc}}(\varepsilon)$ is the discrete program of Theorem 2 throughout, body and appendix; $D^{-}(\varepsilon)$ names a convex relaxation of it, and $D^{-} \le D^{\mathrm{disc}} \le d_{\varepsilon}$. Let $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_F$ be the Frobenius inner product and norm, let $\mathbf{1}$ be the all-ones matrix and let $\tilde{y} = y - \bar{y}\mathbf{1}$, so that $\mathrm{corr}(y,p) = \langle \tilde{y}, \tilde{p} \rangle / (\|\tilde{y}\|_F \|\tilde{p}\|_F)$ for

nonconstant y and $S = \|\tilde{p}\|_F^2 > 0$. Call i a *leader* of task t when $p_{it} = p_{(1)t}$ and a *runner-up* when $p_{it} = p_{(2)t}$, and let SS_t and m_t be the centred sum of squares and the mean of column t, so that $S = \sum_t SS_t + K\sum_t (m_t - \bar{p})^2$. Write $\ell_t(a) = p_{(1)t} - p_{a(t)t}$, and let $d_t(i)$ be the squared distance from $p_{\cdot t}$ to $\{v \in \mathbb{R}^K : v_i \geq v_j \ \forall j\}$. Since $\mathscr{C}_a$ is the product of those per-column cones, $d(a) = \sum_t d_t(a(t))$. Every choice picks some policy, so $\ell_t(a) \in \{0\} \cup [g_t, L_t]$ and $R(a) \leq \frac{1}{T}\sum_t L_t$, and the d_ε problem is feasible exactly when $\sum_t L_t > T\varepsilon$. Otherwise $d_\varepsilon = +\infty$ and $r_\varepsilon = 0$, no correlation requirement being needed. Arguments below that use rationality of the board say where they do. Every floor this paper prints on a probability is read off one-sided 95% Clopper-Pearson lower ends, Bonferroni over the family its sentence names. A mean, a count or an expectation built from several such probabilities is bounded by evaluating it at their lower ends, and the step floor of Fig. 3 inverts a noncentral t instead. The matched null's floors pool its five streams' draws, so they bound the probability under the five calibrated scales taken together, for which stratified sampling makes the pooled binomial interval conservative. The scan of the three structured laws and the random-board null were enlarged after first inspection, as `code/PREREG_margins.md` records. In prose a range is rounded outwards, except a range of crossings. A bound, an error, a p-value, a bar or a crossing is rounded in the direction that weakens its claim. Every other value, a certificate included, is rounded to the nearest place.

Ties and tie-breaking. Ties among an evaluator's own maxima are resolved against it throughout, including in the agreement count; the optimistic rule would raise Ctrl-World's success board from $6/7$ to $7/7$, certifying a tolerance the evaluator violates.

A.1 Proofs of the statements in the body

Proposition 1 (correlation and cone distance)**.** *Let $K \geq 2$, let $S > 0$ and let a be a choice. Then* $\max\{\mathrm{corr}(y,p) : y \in \mathscr{C}_a,\ y \text{ nonconstant}\} \leq \sqrt{1 - d(a)/S}$*, the maximum being attained. If $d(a) < S$ this is an equality, attained at the projection $\Pi_{\mathscr{C}_a}(p)$. If $d(a) = S$ and $T \geq 2$ it is still an equality, both sides being zero and the maximum attained at a column-constant table. If $d(a) = S$ and $T = 1$ the left side can be strictly negative and the identity fails, so only the inequality survives in general.*

Proof of Proposition 1. Cut out by the inequalities $y_{it} - y_{a(t)t} \leq 0$, each involving two entries of one column, $\mathscr{C}_a$ is a closed convex cone containing every column-constant matrix and its negative. So $\mathscr{C}_a = \tilde{\mathscr{C}} \oplus \mathrm{span}(\mathbf{1})$ for $\tilde{\mathscr{C}} = \mathscr{C}_a \cap \mathbf{1}^\perp$, and splitting the squared distance along that sum gives $d(a) = \mathrm{dist}(\tilde{p}, \tilde{\mathscr{C}})^2$. As $\mathrm{corr}(y,p)$ depends on y only through the ray of $\tilde{y}$, and y is nonconstant exactly when $\tilde{y} \neq 0$, the supremum equals $\max\{\langle u, \tilde{p}\rangle/\sqrt{S} : u \in \tilde{\mathscr{C}}, \|u\|_F = 1\}$, attained because that set is compact and, by $K \geq 2$, nonempty. For $q = \Pi_{\tilde{\mathscr{C}}}(\tilde{p})$, Moreau's inequality (Moreau, 1962) $\langle \tilde{p} - q, u\rangle \leq 0$ and that of Cauchy and Schwarz give $\langle u, \tilde{p}\rangle \leq \langle u, q\rangle \leq \|q\|_F$ on unit vectors, with equality at $u = q/\|q\|_F$ when $q \neq 0$, while Pythagoras gives $\|q\|_F^2 = S - d(a)$. The maximum is therefore $\sqrt{1 - d(a)/S}$ when $q \neq 0$, that is when $d(a) < S$, and at most zero when $q = 0$, since then $\langle u, \tilde{p}\rangle \leq 0$ throughout $\tilde{\mathscr{C}}$. At $d(a) = S$ the constant vector at the column mean, lying in every per-column cone, gives $d_t(i) \leq SS_t$ and $d(a) \leq \sum_t SS_t \leq S$, so equality forces $m_t = \bar{p}$ for every t. A column-constant b with $b_{it} = \beta_t$ then has $\tilde{b} \in \tilde{\mathscr{C}}$ and $\langle \tilde{b}, \tilde{p}\rangle = K\sum_t (\beta_t - \bar{\beta})(m_t - \bar{p}) = 0$, so for $T \geq 2$ a nonconstant β makes the maximum exactly zero. For $T = 1$ take $K = 2$, $p = (1,0)^\top$ and $a(1) = 2$, so $S = d(a) = \frac{1}{2}$; every y in that cone has $\tilde{y} = (-s/2, s/2)$ with $s \geq 0$, hence correlation -1 when nonconstant, against a right side of zero.

If y is nonconstant with $\mathrm{corr}(y,p) > r_\varepsilon$ and a selects a taskwise maximiser of y, then $y \in \mathscr{C}_a$, so $R(a) > \varepsilon$ would force $\mathrm{corr}(y,p) \leq r_\varepsilon$; hence $R(a) \leq \varepsilon$ under every tie rule, and so $\varepsilon^\star(r) \leq \varepsilon$ for every $r > r_\varepsilon$; the converse holds except in the degenerate case $d_\varepsilon = +\infty$, where $r_\varepsilon = 0$ and already $\varepsilon^\star(0) \leq \varepsilon$. Only the inequality is used there, so the failure at $d(a) = S$ leaves the certificate valid and merely conservative. And $\mathscr{C}_a$ is the *weak* argmax cone, the strict-argmax set being its relative interior in each column, so a strict-argmax evaluator attains $\sqrt{1 - d(a)/S}$ only as a supremum unless $\Pi_{\mathscr{C}_a}(p)$ has a strict maximum throughout. For the demotions realising Theorem 1 it never has, the projection pooling each demoted policy with its leader, so the certificate is stated with the strict inequality $\mathrm{corr}(y,p) > r_\varepsilon$. □

Proof of Lemma 1. Fix a task t, write $v = p_{\cdot t}$ and $M = p_{(1)t}$, fix i and put $\ell = M - v_i$. The pooled level of the lemma is $c^\star = M - \ell/2$ when $\ell \leq 2g_t$ and $c^\star = M - (\ell + g_t)/3$ when $\ell > 2g_t$.

Minimise $\|u-v\|^2$ over the cone by first fixing the level $c=u_i$: coordinate i then contributes $(c-v_i)^2$, while each $j\neq i$ is free subject to $u_j\le c$ alone and so contributes $((v_j-c)_+)^2$. This gives the variational form

$$d_t(i)=\min_{c\in\mathbb{R}}\varphi_i(c),\qquad \varphi_i(c)=(c-v_i)^2+\sum_{j\neq i}\big((v_j-c)_+\big)^2, \tag{8}$$

from which all three claims follow. Each summand is convex and the first is coercive, so the minimum is attained at some $c^\star$, and differentiating gives $c^\star-v_i=\sum_{j\neq i}(v_j-c^\star)_+\ge 0$.

The floor. Assume $\ell>0$, the case $\ell=0$ being trivial. Every summand of (8) is nonnegative, so keeping only the two indices the published profile names, a leader $j^\star\neq i$ at M and, when $\ell>g_t$, a runner-up $j^{\star\star}\notin\{i,j^\star\}$ at $M-g_t$, leaves a lower bound on φ_i and hence on $d_t(i)$. Translating M to 0, that bound is $\min_c(c+\ell)^2+((-c)_+)^2+((-g_t-c)_+)^2$, which is (8) for the three-entry column $(0,-g_t,-\ell)$. When $\ell\le g_t$ the runner-up need not be distinct from i and only the leader's term survives, but there $\ell\le 2g_t$ and the calculation below returns the leader-only value anyway, so the bound is valid at every ℓ.

That minimand is convex and piecewise quadratic with kinks at $c=0$ and $c=-g_t$, so solving its stationarity condition on each piece and keeping the root lying in that piece gives the value outright. On $c\ge 0$ both hinges vanish and $(c+\ell)^2$ is least at $c=0$, worth ℓ^2. On $-g_t\le c\le 0$ only the leader is active, the derivative is $2(c+\ell)+2c$ and $c=-\ell/2$, which lies in the piece exactly when $\ell\le 2g_t$ and is worth $\ell^2/2$. On $c\le -g_t$ both are active, the derivative is $2(c+\ell)+2c+2(c+g_t)$ and $c=-(\ell+g_t)/3$, which lies in the piece exactly when $\ell\ge 2g_t$ and is worth $\frac{2}{3}(\ell^2-\ell g_t+g_t^2)=\ell^2/2+(\ell-2g_t)^2/6$. The last two pieces partition $[0,\infty)$ and agree at $\ell=2g_t$, which is h_t.

Its shape. $h_t'(\ell)$ is ℓ below the kink and $(4\ell-2g_t)/3$ above, both nonnegative and both equal to $2g_t$ at $\ell=2g_t$, and $h_t''\in\{1,4/3\}$, so h_t is increasing, convex and C^1, and it exceeds $\ell^2/2$ exactly above $2g_t$. It reads the column only through g_t and ℓ.

What is given up. Below the runner-up, nothing: on $\ell\le 2g_t$ the leader-only minimiser $c=-\ell/2$ already lies at or above $-g_t$, so the runner-up's hinge is inactive there and naming the runner-up cannot move the minimiser or the value. Above it, the unpublished tail: the bound drops $K-3$ nonnegative hinges of the same φ_i, and is tight exactly when $p_{(3)t}$ lies at or below the pooled level.

Equality. $c^\star$ as stated is the minimiser of the two-index bound, and by the displayed stationarity it is the minimiser of φ_i itself precisely when no further entry is active there, that is when every entry outside $\{i,j^\star,j^{\star\star}\}$ is at or below $c^\star$; a further active entry adds a strictly positive term at $c^\star$ and strictly raises the minimum, both minimands being strictly convex. On $\ell\le 2g_t$ this is the midpoint criterion, $c^\star=(v_i+M)/2$, and the runner-up satisfies it automatically there, $-g_t\le -\ell/2$.

The runner-up. Let i be a runner-up, the case $g_t=0$ being trivial since i is then a leader. If $g_t>0$ the leader is unique, $\ell=g_t\le 2g_t$ puts $h_t(g_t)=g_t^2/2$, and every entry outside $\{i,j^\star\}$ is at most $v_i=p_{(2)t}<(v_i+M)/2=c^\star$, so the criterion holds and $d_t(i)=g_t^2/2$ with no further hypothesis. Finally $d_t(i)=0$ exactly when i is a leader, since v itself then lies in the cone. □

Proof of Theorem 1. Let $\varepsilon\ge 0$ and assume $\sum_t L_t>T\varepsilon$. Write $\mathscr{U}(\varepsilon)$ for the disjoint pairs (U,B) of task sets with $\sum_{t\in U}g_t+\sum_{t\in B}L_t>T\varepsilon$; the hypothesis says exactly that $(\emptyset,[T])$ lies in it, so (4) is a minimum over a nonempty finite family.

The bottom rung. Each hinge $((v_j-c)_+)^2$ of (8) is nondecreasing in v_j, so $d_t(i)$ is nondecreasing in every entry other than v_i. Let i be a bottom policy of task t, so $v_i=M-L_t$ with $M=p_{(1)t}$. Every entry other than v_i and one leader is at most the runner-up $M-g_t$, so raising those $K-2$ middle entries to it gives $d_t(i)\le u_t$. The raised column's φ_i is least at $c=M-L_t/2$, pooling i with the leader alone, when $L_t\le 2g_t$, and at the mean $M-((K-2)g_t+L_t)/K$ of the whole column otherwise, a mean below $M-g_t$ exactly when $L_t>2g_t$; the two costs are those of the statement.

The bound. Given $(U,B)\in\mathscr{U}(\varepsilon)$, let a take a runner-up on U, a bottom policy on B and a leader elsewhere. Then $R(a)>\varepsilon$, and by Lemma 1's runner-up identity and the bottom rung $d(a)\le\frac{1}{2}\sum_{t\in U}g_t^2+\sum_{t\in B}u_t$; minimising gives $d_\varepsilon\le D^+(\varepsilon)$. The column-mean table lies in every per-column cone, so $d_\varepsilon\le\sum_t SS_t\le S$, while $D^+(\varepsilon)$ can exceed S, which is why the endpoint is read as zero there:

on the single column $(1, 1, \frac{1}{2}, \frac{1}{2}, 0)$ the only rung of (4) that buys a budget below 1 is $u_1 = \frac{4}{5}$, against $S = \frac{7}{10}$.

Equality. $D^+(\varepsilon) = d_\varepsilon$ if and only if some minimiser $a^\star$ of the d_ε problem demotes each task it concedes to a runner-up or to a bottom policy i with $d_t(i) = u_t$. If it does, its runner-up and bottom tasks form a pair in $\mathscr{U}(\varepsilon)$ of objective $d(a^\star) = d_\varepsilon$; conversely a minimising pair gives a choice a with $d_\varepsilon \le d(a) \le D^+(\varepsilon) = d_\varepsilon$, whose bottom demotions must then cost exactly u_t. When $L_t \le 2g_t$ every bottom policy costs u_t: the raised and the true φ_i coincide on $[M - g_t, \infty)$, which holds the raised minimiser, and both are continuously differentiable. When $L_t > 2g_t$ the raised minimiser lies below $M - g_t$, where a middle entry below the runner-up strictly lowers φ_i, so a bottom policy costs u_t exactly when every middle entry sits at the runner-up. With $K \le 3$ every middle entry is a runner-up and every policy is a leader, a runner-up or a bottom one, so every minimiser qualifies. From four policies on it can fail, and $D^+(\varepsilon)/d_\varepsilon$ is unbounded. On n tasks with column $(1, 1, 1 - 1/n, 0)$ at a budget $T\varepsilon = 1 - 1/(2n)$ a runner-up demotion buys nothing, so (4) pays one bottom demotion at $u_t = \frac{3}{4}$, while the cheapest costly choice takes the third policy on every task, at $\frac{2}{3}n^{-2}$ each. The ratio is $9n/8$, which is 144 at $n = 128$. That looseness is the profile's, not the program's. Raising every middle entry of p to its runner-up leaves $(g_t)_t$, $(L_t)_t$ and K unchanged and makes every policy a leader, a runner-up or a bottom one of cone cost u_t, so on the raised board $d_\varepsilon = D^+(\varepsilon)$, and no upper bound on d_ε that reads only those numbers is smaller. The equality argument above, with $B = \emptyset$, shows that the runner-up program $D_g^+(\varepsilon)$ of Sec. 3, where it exists, is d_ε exactly when some minimiser concedes only runner-ups.

Without the hypothesis no choice is costly, $d_\varepsilon = +\infty$ and $r_\varepsilon = 0$, and the ranges, not the margins, decide which side of that line a board is on. On

$$p = \frac{1}{10}\begin{bmatrix} 3 & 7 & 2 \\ 3 & 7 & 2 \\ 1 & 0 & 2 \end{bmatrix}$$

every leader ties its runner-up, so no runner-up demotion buys anything; yet $\sum_t L_t = \frac{9}{10}$, and at $\varepsilon = .01$ the first task's bottom rung $u_1 = \frac{2}{75}$ is d_ε itself, three policies making the endpoint exact: $d_\varepsilon/S = \frac{1}{18}$ and $r_\varepsilon = \sqrt{17/18}$. □

Proof of Theorem 2. Let $\varepsilon \ge 0$ and assume $\sum_t L_t > T\varepsilon$. The hypothesis makes the constraint set nonempty, $\ell = L$ being feasible. Let a satisfy $R(a) > \varepsilon$ and put $\ell_t = \ell_t(a)$; each ℓ_t lies in $\{0\} \cup [g_t, L_t]$ and $\sum_t \ell_t = TR(a) > T\varepsilon$, so ℓ is feasible for that program, and Lemma 1 gives $d(a) = \sum_t d_t(a(t)) \ge \sum_t h_t(\ell_t) \ge D^{\mathrm{disc}}(\varepsilon)$. Minimising over such a gives $d_\varepsilon \ge D^{\mathrm{disc}}(\varepsilon)$, which is the claim. Since $h_t(\ell) \ge \ell^2/2$, the half-square program below is at most $D^{\mathrm{disc}}(\varepsilon)$, and the envelope step below puts $D^-(\varepsilon)$ at most that program, so the weaker endpoint follows with it, and $D^{\mathrm{disc}}(\varepsilon) \le d_\varepsilon \le S$ makes the square root well defined. Nothing in (5) reads a cell of the board: h_t is a function of (g_t, ℓ) and its feasible set of (g_t, L_t), which is what Cor. 1 asserts. Printing more of each column is priced in App. A.3.

Sharpness. The two charges part only above $2g_t$, so the floor is tighter exactly on the boards where buying the budget forces a demotion deeper than twice a margin: in Table 2, Ctrl-World's two, AgentRewardBench and WMT23.

No larger bound from the same numbers. From four policies on, no lower bound on d_ε that reads only $(g_t)_t$, $(L_t)_t$ and K exceeds $D^{\mathrm{disc}}(\varepsilon)$, unless every minimiser of (5) concedes each of its tasks at its full range and meets the budget exactly. Take any other minimiser. On a board with the same margins and ranges, put every entry below the runner-up at $p_{(1)t} - L_t$, except one policy on each conceded task, placed at the loss the minimiser assigns; four policies leave an entry at $p_{(1)t} - L_t$ beside it, so the range is kept. Every entry but that policy, a leader and a runner-up lies at or below the pooled level of Lemma 1, so the policy's cone cost is $h_t(\ell_t)$, and taking it on each conceded task costs $D^{\mathrm{disc}}(\varepsilon)$. That choice is costly where the minimiser exceeds the budget. Where it meets the budget exactly, deepening one loss short of its range by $\eta > 0$ keeps the margins and ranges and gives a costly choice within $O(\eta)$ of $D^{\mathrm{disc}}(\varepsilon)$. Theorem 2 holds d_ε at or above the endpoint on every board, so on these boards d_ε attains or approaches it, and no bound reading only those numbers can exceed it. With the raised board above, neither program can be tightened from those numbers outside the excepted case. In that case a costly choice must concede more than the minimiser can, and the endpoint can sit strictly below d_ε on every board sharing them.

Water-filling on the marginal price. The program $\frac{1}{2}\min\{\sum_t \ell_t^2 : \ell_t \in \{0\}\cup[g_t, L_t],\ \sum_t \ell_t \geq T\varepsilon\}$, which $D^{\mathrm{disc}}(\varepsilon)$ dominates because $h_t(\ell) \geq \ell^2/2$, is not convex either, its feasible set omitting $(0, g_t)$ on each task. Replace $\ell \mapsto \ell^2$ on $\{0\}\cup[g_t, L_t]$ by its convex envelope on $[0, L_t]$, namely the chord $f_t(\ell) = g_t\ell$ for $\ell \leq g_t$ and $f_t(\ell) = \ell^2$ above, and put $D^-(\varepsilon) = \frac{1}{2}\min\{\sum_t f_t(\ell_t) : 0 \leq \ell_t \leq L_t,\ \sum_t \ell_t \geq T\varepsilon\}$. Since f_t agrees with ℓ^2 on the discrete loss set, which the box $[0, L_t]$ contains, $D^-(\varepsilon)$ is at most the half-square program, and so $D^-(\varepsilon) \leq D^{\mathrm{disc}}(\varepsilon) \leq d_\varepsilon$, so $\sqrt{1 - D^{\mathrm{disc}}(\varepsilon)/S}$ is the sufficient endpoint Theorem 2 publishes and $\sqrt{1 - D^-(\varepsilon)/S}$ a valid but more conservative one. It dominates the plain relaxation to $[0, L_t]$, which keeps ℓ^2 throughout and so charges almost nothing for the first g_t units of loss that no choice can buy, where the chord charges $g_t\ell \geq \ell^2$. The two values can differ: for $K = T = 2$ and

$$p = \begin{bmatrix} 1 & .1 \\ 0 & 0 \end{bmatrix}$$

at $\varepsilon = \frac{1}{4}$, $D^- = \frac{41}{200}$ while $D^{\mathrm{disc}} = d_\varepsilon = \frac{1}{2}$.

Assume $\varepsilon > 0$, the case $\varepsilon = 0$ having optimum $\ell = 0$, fix a price $\lambda \geq 0$, and minimise $f_t(\ell) - \lambda\ell$ over $[0, L_t]$ one task at a time. The chord has slope $g_t - \lambda$ and the quadratic piece has its vertex at $\lambda/2$, so the minimisers are $\ell = 0$ for $\lambda < g_t$, the whole interval $[0, g_t]$ at $\lambda = g_t$, $\ell = g_t$ for $g_t < \lambda \leq 2g_t$, $\ell = \lambda/2$ for $2g_t \leq \lambda \leq 2L_t$, and $\ell = L_t$ above, the overlapping descriptions agreeing at $\lambda = 2g_t$ and $\lambda = 2L_t$. Let $\overline{\ell}_t(\lambda)$ and $\underline{\ell}_t(\lambda)$ be the largest and the smallest minimiser, which differ only at $\lambda = g_t$, where the task is exactly indifferent along its chord. Both are nondecreasing, and $\sum_t \overline{\ell}_t(2\max_t L_t) = \sum_t L_t > T\varepsilon$, so the prices with $\sum_t \underline{\ell}_t(\lambda) \leq T\varepsilon \leq \sum_t \overline{\ell}_t(\lambda)$ form a nonempty compact interval in $[0, 2\max_t L_t]$; let λ be its least element. Each per-task minimiser set being an interval, some $\ell^\star$ has $\underline{\ell}_t(\lambda) \leq \ell_t^\star \leq \overline{\ell}_t(\lambda)$ and $\sum_t \ell_t^\star = T\varepsilon$, so at a price where a task is indifferent the optimum interpolates along that task's chord. Any such $\ell^\star$ minimises the Lagrangian and is feasible with the constraint active, so it solves the convex program, and $2D^-(\varepsilon) = \sum_t f_t(\underline{\ell}_t(\lambda)) + \lambda(T\varepsilon - \sum_t \underline{\ell}_t(\lambda))$, because $f_t(\ell) - \lambda\ell$ is constant along each minimiser interval. The strict hypothesis $\sum_t L_t > T\varepsilon$ is what bounds that interval of prices and pins the optimal value: at the boundary $\sum_t L_t = T\varepsilon$ the only feasible point is $\ell = L$ and every $\lambda \geq 2\max_t L_t$ is admissible, though there the bound is vacuous anyway. □

Proof of Corollary 1. This is Theorem 1 and Theorem 2 together under their common hypothesis, with the first's equality when $K \leq 3$; neither program reads more than $(g_t)_t$, $(L_t)_t$ and K. □

Proof of Proposition 2. Write $\mathscr{U}_g(\varepsilon) = \{U \subseteq [T] : \sum_{t\in U} g_t > T\varepsilon\}$ for the feasible sets of the runner-up program, which the hypothesis makes nonempty. The left bound on $m^\star$ is feasibility. For the right one, if some $t_0 \in U^\star$ had $m^\star - g_{t_0} > T\varepsilon$ then $U^\star \setminus \{t_0\}$ would lie in $\mathscr{U}_g(\varepsilon)$, against inclusion-minimality. The exact dynamic programme returns a cost-minimiser with every margin positive, which is inclusion-minimal, deleting a member while keeping feasibility being a strict cost saving. For (6), whose middle term is $D_g^+(\varepsilon)/S$, $2D_g^+(\varepsilon) = \sum_{t\in U^\star} g_t^2 = \mu(\varepsilon)m^\star$ and $S = KT\sigma^2$, so $D_g^+(\varepsilon)/S = \mu(\varepsilon)m^\star/(2KT\sigma^2)$, and substituting the two bounds on $m^\star$ cancels one factor of T. Here $\mu(\varepsilon) > 0$, since $m^\star > T\varepsilon \geq 0$ forces a positive margin in $U^\star$; the left inequality is strict and the right one is attained. The relative width is $\min_{t\in U^\star} g_t/(T\varepsilon) \leq 1/(T\varepsilon)$, so the number of tasks appears only there, and the law pins the squared deficit $1 - r_\varepsilon^2$ rather than r_ε. Both $\mu(\varepsilon)$ and σ^2 remain statistics of the board until a sampling model is imposed.

The limit. Let the columns be independent draws from a fixed law on $[0,1]^K$ whose induced top-two margin g has a density, and fix ε with $0 < \varepsilon < \mathbb{E}[g]$. Both are needed: otherwise the runner-up program's hypothesis fails almost surely for large T, or the truncation level below is not pinned. Let Q solve $\mathbb{E}[g\mathbb{I}\{g \leq Q\}] = \varepsilon$ with $\mathbb{I}$ the indicator, which exists because that map is continuous and runs from 0 to $\mathbb{E}[g] > \varepsilon$; where several Q solve it, g has no mass between them, so $\mathbb{E}[g^2\mathbb{I}\{g \leq Q\}]$ is the same at each. Then, almost surely as $T \to \infty$, $m^\star/T \to \varepsilon$ and $\mu(\varepsilon) \to \mathbb{E}[g^2 \mid g \leq Q]/\mathbb{E}[g \mid g \leq Q]$, while σ^2 converges to the variance of a cell drawn by sampling a column and then a row uniformly, so $D_g^+(\varepsilon)/S$ and $\sqrt{1 - D_g^+(\varepsilon)/S}$ have limits depending only on the column law and on ε.

Let $\mathrm{LP}(\tau) = \min\{\sum_t x_t g_t^2 : 0 \leq x_t \leq 1, \sum_t x_t g_t \geq \tau\}$ relax the knapsack; a unit of coverage from task t costs g_t, so an exchange argument on its single constraint makes filling in nondecreasing order of g_t

optimal. Each $U \in \mathscr{U}_g(\varepsilon)$ gives a feasible $x = e_U$, its indicator vector, so $\mathrm{LP}(T\varepsilon) \le 2D_g^+(\varepsilon)$, while the fill prefix, extended by one further task when its mass lands exactly on $T\varepsilon$, is a feasible U of cost at most $\mathrm{LP}(T\varepsilon) + 2\max_t g_t^2$. On a rational board that extension is unnecessary and the constant halves, the strict constraint $\sum_{t\in U} g_t > T\varepsilon$ being equivalent to $\sum_{t\in U} g_t \ge M$ for M the least integer above the scaled target; the substitution is legitimate only because the scaled margins are integers. The empirical partial moments $\frac{1}{T}\sum_t g_t \mathbb{I}\{g_t \le x\}$ and $\frac{1}{T}\sum_t g_t^2 \mathbb{I}\{g_t \le x\}$ converge uniformly in x, almost surely, to $\mathbb{E}[g\mathbb{I}\{g \le x\}]$ and $\mathbb{E}[g^2\mathbb{I}\{g \le x\}]$. The first limit being continuous and nondecreasing, the level at which the fill stops converges to the set of solutions, on which the second limit is constant, and $\mathrm{LP}(T\varepsilon)/T \to \mathbb{E}[g^2\mathbb{I}\{g \le Q\}]$. The overshoot bound gives $m^\star/T \in (\varepsilon, \varepsilon + 1/T]$, so $\mu(\varepsilon) = 2D_g^+(\varepsilon)/m^\star \to \mathbb{E}[g^2\mathbb{I}\{g \le Q\}]/\varepsilon$, the stated ratio because $\mathbb{E}[g\mathbb{I}\{g \le Q\}] = \varepsilon$. Finally $\sigma^2 = S/(KT)$ averages KT terms $(p_{it} - \bar{p})^2$ whose column averages obey the strong law. The envelope endpoint converges by the machinery of Prop. 8, proved in App. A.3 without this proposition. Applied to f_t in place of the rung envelope $\hat{\varphi}_t$ used there, that machinery bounds the program $D^-(\varepsilon)$ and not $D^{\mathrm{disc}}(\varepsilon)$: f_t's level $\overline{\ell}_t(\lambda)$ is at least g_t once $\lambda > g_t$, so the optimal price is at most 1 once the margin mass clears the budget, and $\lambda \mapsto \min_\ell(f_t(\ell) - \lambda\ell)$ is bounded and 1-Lipschitz on $[0,1]$, which is all the uniform law of large numbers there needs. So D_g^+/S and D^-/S are both free of T. That cancellation is the reason the limit exists, and not a coincidence. The regret target is a per-task mean, so the loss mass a violating choice must buy grows like $T\varepsilon$, its cone cost grows with it, and $S = KT\sigma^2$ is extensive for the same reason. Both being extensive, their ratio is intensive, so doubling a benchmark only shrinks the $O(1/(T\varepsilon))$ overshoot. Proposition 2's closed form lies strictly above the runner-up program's endpoint $\sqrt{1 - D_g^+(\varepsilon)/S}$, by the left inequality of (6). It lies above the bar on WorldGym and below Theorem 1's endpoint on Ctrl-World's instruction board, so it bounds neither side of the bracket on r_ε, and licenses neither conclusion. □

Proof of Proposition 3. Let y satisfy $\mathrm{MMRV}(y,p) \le v$ and let a take a taskwise maximiser of y, any one of them where y ties. Fix a task t, write $i = a(t)$ and $\ell = p_{(1)t} - p_{it}$, and let $A = \{j : p_{jt} > p_{it}\}$. If $\ell = 0$ then A is empty and the task contributes nothing to either side, so assume $\ell > 0$; then A is nonempty and contains a leader at distance ℓ from p_{it}.

Because y selects i on task t we have $y_{it} \ge y_{jt}$ for every j, so for $j \in A$ the board's strict ordering $p_{jt} > p_{it}$ is not matched by the scores: $[y_{jt} > y_{it}] = 0$ while $[p_{jt} > p_{it}] = 1$. The pair is therefore violated *at* j, whether the scores separate the two or tie them, and the summand of (7) at j is at least $p_{jt} - p_{it}$. Summing those summands over $j \in A$, the task contributes at least $c_t^-(i) = \sum_{j\in A}(p_{jt} - p_{it}) \ge \ell$ to $KT\,\mathrm{MMRV}$. It is the violation at j, not at i, that carries this bound: if y ties i with every member of A then $[y_{it} > y_{jt}] = 0 = [p_{it} > p_{jt}]$ and the summand at i is zero. Summing over tasks gives $KTv \ge \sum_t c_t^-(a(t)) \ge \sum_t \ell_t = TR(a)$, which is part (i).

For part (ii) assume in addition that $y_{it} > y_{jt}$ for every $j \ne i$ and every task. Each $j \in A$ then also makes $[y_{it} > y_{jt}] = 1 \ne 0 = [p_{it} > p_{jt}]$, so the pair is violated at i as well and the summand at i is at least $\max_{j\in A}(p_{jt} - p_{it}) = \ell$. The task's charge is then $c_t^+(i) = \ell + c_t^-(i) \ge 2\ell$, and the same summation gives $R(a) \le \frac{K}{2}v$. The strictness hypothesis cannot be dropped: on the board p with the score column y,

$$p = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad y = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix},$$

y has $\mathrm{MMRV} = 1/2$ and, under a tie resolved against the evaluator, regret 1, which exceeds $\frac{K}{2}v = 1/2$ and meets $Kv = 1$.

In both cases the worst regret compatible with a published v is the maximum of $R(a)$ over choices whose total charge is at most KTv, the charge being $c_t^-(a(t))$ in case (i) and $c_t^+(a(t))$ in case (ii). Each charge is attained, so neither constraint set can be tightened: in case (ii) by the column that puts $a(t)$ strictly above every rival and otherwise preserves the board's order, and in case (i) by the column that ties $a(t)$ with every policy above it and preserves the order below. The maximum is the knapsack of the statement, solved exactly over integer charges once the board is rational. Equality with the closed form requires every conceded task to be conceded to its runner-up, where A is a single leader and the charge is exactly the loss in case (i) and twice it in case (ii). Neither factor can be lowered: the two-policy board above attains part (i). On a board conceding the same gap on every task, an evaluator that puts the runner-up strictly first is charged that gap twice on each task it concedes, so the budget KTv buys exactly $Kv/2$ of regret and part (ii) is met with equality. □

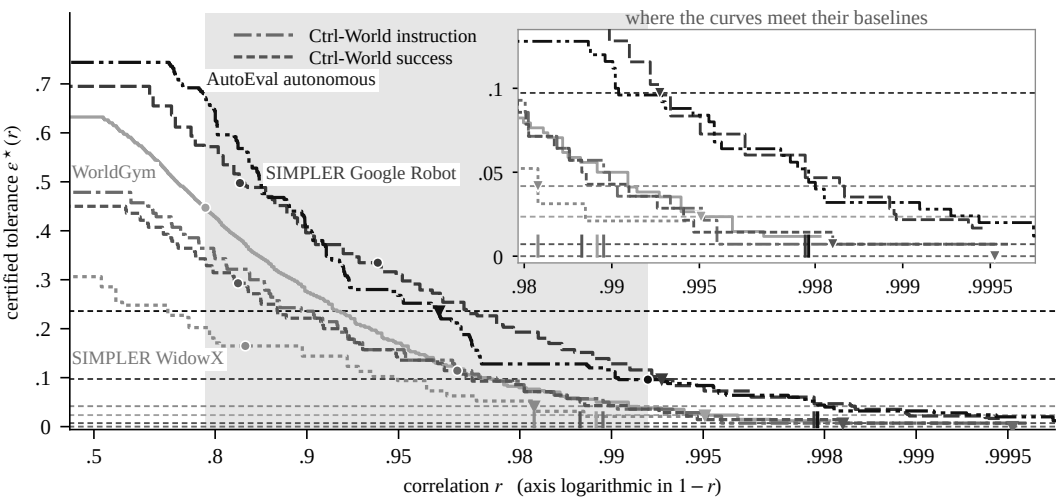


Figure 2: What a published correlation certifies for the per-task decision, on six of the nine distinct boards. Three are left out at this size: one because its .450 would stretch the shaded band across the axis, and two because they lie below .19 and .14 throughout. The shaded band spans these six boards' recomputed correlations. Each step curve is the exact worst selection regret $\varepsilon^\star(r)$, each horizontal dashed line the regret of deploying that board's best fixed policy instead. Circles sit at the recomputed correlations, triangles at the crossings past which a curve certifies at or below its horizontal line, and AutoEval's circle is the only one past its own triangle. Short ticks mark the bar $r_{.05}$ of Table 2, the inset magnifies the region past .98, and both horizontal axes are logarithmic in $1-r$.

Proof of Proposition 4. A choice a leads the board on task t when $p_{a(t)t} = p_{(1)t}$. Put $N = \{t : p_{a(t)t} < p_{(1)t}\}$, so $|N| \le T-m$, with $\ell_t(a) = 0$ off N and $\ell_t(a) \le L_t$ on N; hence $TR(a) = \sum_{t\in N} \ell_t(a) \le \sum_{t\in N} L_t \le \sum_{j\le T-m} L_{[j]}$, since at most $T-m$ ranges sum to at most the $T-m$ widest. Taking a minimiser on those widest tasks and a leader elsewhere attains it. □

A.2 The exchange rate on the corpus, and under a declared accuracy

Crossings. Section 2 reports the correlation past which $\varepsilon^\star(r)$ certifies at or below the regret of the board's best fixed policy, the crossing Fig. 2 marks with a triangle. The lowest are AutoEval's, .9521 on its SIMPLER board and .9632 on its autonomous one. Then come .9821 on SIMPLER WidowX, .9932 on SIMPLER Google Robot and .9951 on WorldGym, and last .9983 on Ctrl-World's success board, .9991 on AgentRewardBench, .9996 on Ctrl-World's instruction board and .9997 on WMT23. Each value is the exact crossing rounded up to four places, so every correlation above it certifies at or below that board's best fixed policy.

Why the two decisions swap places. At $r = 1$ the constraint $d(a) \le S(1-r^2) = 0$ admits only choices that take a maximum on every task, so $\varepsilon^\star(1) = 0$ for the per-task decision. The aggregate certificate at $r = 1$ admits only the policy of highest mean, and its value on the task scale is that policy's own per-task regret, zero exactly when one policy leads every task. Both certificates are step functions of $1-r$. So on any board whose best fixed policy does not lead everywhere, the per-task certificate is zero on a whole neighbourhood of $r = 1$ while the aggregate one on the task scale is not, and the finer decision is the better constrained throughout it. The board decides how far below $r = 1$ that persists, since two step functions can cross more than once. The corpus straddles that crossover, and AgentRewardBench straddles it within one board: its judge, at $r = .965$, falls on one side and its rule-based checker, at $r = .931$, on the other. At the correlation the introduction opens with, the aggregate certificate is zero on three boards, WorldGym's and Ctrl-World's two, and on each of those the per-task certificate exceeds the best fixed policy's regret.

One board, two evaluators. The four tasks AutoEval's SIMPLER comparison covers are a subset of the five its autonomous proxy covers, with identical real cells. Restricted to those four, the autonomous proxy carries $r = .996$ and certifies .060; the simulator carries $r = .450$ and certifies .870, the whole range of the board.

Proposition 5 (a declared per-cell accuracy). *Fix $r \in (0,1]$ and $\delta \geq 0$, write $\delta(y) = \inf\{\|\alpha y + \beta\mathbf{1} - p\|_\infty : \alpha > 0,\ \beta \in \mathbb{R}\}$, and let $\varepsilon^\star(r,c)$ be* (2) *restricted to choices with $\ell_t(a) \leq c$ on every task. Then every y with $\mathrm{corr}(y,p) \geq r$ and $\delta(y) \leq \delta$ induces a choice of regret at most $\varepsilon^\star(r,2\delta)$; and capping instead $\max_t \|\Pi_{\mathscr{C}_a}(p)_{\cdot t} - p_{\cdot t}\|_\infty$ by δ gives a value $\underline{\varepsilon}(r,\delta)$ attained by some y in that class.*

Proof of Proposition 5. The infimum need not be attained; on some tables it is approached only as $\alpha \downarrow 0$. So fix $\eta > 0$ and let $u = \alpha y + \beta\mathbf{1}$ with $\alpha > 0$ and $\|u - p\|_\infty \leq \delta + \eta$. Since $\alpha > 0$, u and y have the same taskwise argmax, so $u_{a(t)t} \geq u_{it}$ for every i, and in particular for a leader i_1 of task t. Then $p_{a(t)t} \geq u_{a(t)t} - \delta - \eta \geq u_{i_1 t} - \delta - \eta \geq p_{i_1 t} - 2\delta - 2\eta$, so $\ell_t(a) \leq 2\delta + 2\eta$ on every task; the losses take finitely many values, so letting $\eta \downarrow 0$ gives $\ell_t(a) \leq 2\delta$. Proposition 1 gives $d(a) \leq S(1-r^2)$, so a is feasible for $\varepsilon^\star(r,2\delta)$. The factor two is tight: an evaluator may lower a leader by δ and raise the policy it selects by δ.

For the second, let $r > 0$, let a satisfy $d(a) \leq S(1-r^2)$, which then forces $d(a) < S$, and put $y = \Pi_{\mathscr{C}_a}(p)$. Then $y \in \mathscr{C}_a$, and Proposition 1 gives $\mathrm{corr}(y,p) = \sqrt{1 - d(a)/S} \geq r$, while $\alpha = 1$, $\beta = 0$ bounds $\delta(y) \leq \|y - p\|_\infty$, which is the capped quantity. Ties are resolved against the evaluator as everywhere here, so y is credited with a. The pooling of Lemma 1 moves only entries whose values lie between $p_{a(t)t}$ and $p_{(1)t}$, and moves them to a mean in that same interval, so the displacement on task t lies in $[\ell_t(a)/2, \ell_t(a)]$, and equals $g_t/2$ at a runner-up. A displacement cap at δ therefore excludes every rung a loss cap at 2δ excludes, and more on the deeper demotions, which is why the two caps bracket the class rather than coincide.

Both values are (2) computed over the same per-task rungs with those above the cap deleted, so both are exact, and both are nondecreasing in δ. The accuracy a board must declare to certify a tolerance ε is any δ with $\varepsilon^\star(r,2\delta) \leq \varepsilon$; the set of such δ is the half-open interval $[0, c/2)$ with c the smallest admitted loss at which $\varepsilon^\star(r,c)$ exceeds ε, so the threshold is that supremum $c/2$ and does not itself certify. On the corpus at $\varepsilon = .05$ it runs from .0325 on WMT23 to .15 on WorldGym, between 1.2 and 6.7 times finer than the calibrated residuals of Sec. 2. □

A declared accuracy converts with no board. The exchange rates above need the whole rung structure, so a benchmark that prints no table cannot evaluate them. Its margin and range profile is enough for a weaker form of the same statement, and that profile is what Sec. 4 already asks for. A calibrated residual at most δ forces $\ell_t(a) \leq 2\delta$ on every task by the argument above, while every loss lies in $\{0\} \cup [g_t, L_t]$, so a task can be conceded only where $g_t \leq 2\delta$ and then by at most $\min(L_t, 2\delta)$. Hence

$$R(a) \ \leq\ \tfrac{1}{T} \sum_{t:\, g_t \leq 2\delta} \min(L_t, 2\delta),$$

attained on a board whose every qualifying task carries an entry at exactly that loss. On the corpus it is below the trivial bound on all eleven pairs, and on three of them it is below what the correlation certifies: .128 against .132 on WMT23, .170 against .189 on AgentRewardBench's rule-based checker and .861 against .870 on AutoEval's SIMPLER bridge. So the declared accuracy of Sec. 4 converts without the board too. With the count's bound and the rank statistic's closed form, the least of the three is tighter than the correlation with the board on ten of eleven pairs, and at or below the best fixed policy's regret on four to its one. This analysis is not registered.

A.3 The bar on the corpus, on random boards and as a board grows

Why the recommendation stops at two numbers per task. A benchmark could print the top j order statistics of each column beside its range. The cheapest costly choice that demotes each task only to a printed rank or to its bottom policy, the latter priced at the column with every unprinted entry raised to the deepest printed one, then bounds d_ε from above. The program that charges any deeper loss by the floor the printed head implies drops only nonnegative terms of the cone cost, so it bounds d_ε from below. At $j = 2$ the two are the sides of Cor. 1, and from $j = K - 1$ the printed numbers are the whole column. On this corpus the two meet on six of the nine boards at two numbers per task and on eight at three. Of the two that a third number adds, AgentRewardBench has four policies, where three numbers are the whole column; with part of each column withheld, a third number closes one board, SIMPLER's Google Robot, and a fourth closes no more. On 1200 random boards with $K \in \{3,\dots,7\}$ and $T \in \{3,\dots,12\}$, each priced at one of four tolerances, three numbers still withhold part of each column on the 729 with five or more policies. There the necessary endpoint

Table 2: The bar at a five-point tolerance, bracketed from the published margins and ranges. The necessary endpoint comes from Thm. 1 and the sufficient endpoint from Thm. 2, except on the four boards with three policies, where Thm. 1 is exact and supplies both (Cor. 1). Both are rounded outwards, so the printed bracket is no narrower than the exact one, and the width column carries the exact difference, rounded up. The width is exactly zero on six of the nine boards, not merely small: the four with three policies, and AutoEval's two, where the two endpoints are the same rational; outward rounding alone separates their printed endpoints. The exact bar is rounded up, as in Table 1. The closed form is Prop. 2's, for the runner-up program D_g^+ rather than Thm. 1's, with $\mu(.05)$ its margin functional, read off the minimiser that program returns, and σ^2 the board's per-cell variance. These three are rounded to the nearest place, and the first two are n/a on WMT23, whose total top-two margin falls below the regret budget. Two boards carry two evaluators each, so nine boards hold the eleven pairs of Table 1; AutoEval's two evaluators are compared on different task sets, so the smaller is its own board. All eleven correlations of Table 1 lie below their boards' necessary endpoints, so the printed profiles and S alone show each to fall short.

	bracket on $r_{.05}$					board statistics	
board	necessary	exact	sufficient	width	closed form	$\mu(.05)$	σ^2
WorldGym	.9887	.9888	.9888	0	.9895	.255	.102
SIMPLER Google Robot	.9947	.9979	.9983	.0036	.9953	.222	.099
SIMPLER WidowX	.9820	.9821	.9821	0	.9828	.109	.027
Ctrl-World, instruction	.9893	.9894	.9894	0	.9884	.228	.082
Ctrl-World, success	.9873	.9874	.9874	0	.9878	.200	.068
AutoEval, autonomous	.9978	.9979	.9979	0	.9980	.118	.122
AutoEval, SIMPLER	.9978	.9979	.9979	0	.9981	.133	.145
AgentRewardBench	.9901	.9931	.9935	.0034	.9912	.127	.045
WMT23	.9553	.9790	.9810	.0257	n/a	n/a	.004

equals d_ε on 543 at three numbers against 270 at two, and on the rest of those boards the largest ratio seen is 4.85, of the necessary program's value to d_ε. Where the margins are thin, more numbers help slowly: WMT23's necessary endpoint is .9553, .9589, .9600, .9658, .9763 and .9788 at two to seven order statistics per task, and reaches its exact bar, .9790, only at eight, the whole column. No depth short of the whole column is exact on every board, so the recommendation of Sec. 4 stays at two numbers per task, the whole column at three policies.

Reading μ on a short board. On the corpus the lower-tail reading of μ needs care, because where $T \leq 7$ the budget is bought from very few tasks. On the SIMPLER Google board one task above the target supplies all of it, so μ is that task's margin rather than a tail average. The bracket holds regardless.

WMT23 at its own scale. Read at tolerances a translation reader would recognise, WMT23's bar is .9968 at two points of human judgement, .9991 at one and .9997 at half a point, the last two above every board's bar at five points, the highest of which is .9979. At half a point the bracket is $[.9994, .9998]$.

Three board laws with structure. Section 3's random boards draw every cell independently. Three laws that do not (a policy effect, a policy and task effect, and one dominant policy) were drawn beside them. Under each the bar clears .97 on at least .970 of 20000 draws and .95 on at least .995, both the one-sided lower end of the binomial interval taken simultaneously over the three laws, at medians from .989 to .994. The least bar a scan draws is the minimum of a sample and falls as the sample grows, so no floor is read off it. The closed form of Prop. 2 tracks the exact bar to a median of .0004 on the independent scan and of .0006 or better under each of these laws. Its largest error is a sample maximum and not a bound, and grows with the sample for that reason: .0293 on that scan, .0505 on a second independent sample of the same law, and up to .1113 under a law with structure. On the corpus its worst error is .0026.

What the scan discards, and which count the bar follows. The shape scan of Sec. 3 reports 499 of its 500 draws, dropping the one on which the margins total less than the budget, so that Prop. 2's closed form, which the scan's error figures read, does not exist. The bracket contains that draw's exact bar too, .977, so the share of draws above .97 is .998 over all five hundred and the filter hides nothing. Over the reported draws the bracket closes on all 123 with three policies. On that one scan's

retained boards the bar correlates with the policy count at .66 and with the task count at .32; once the closed form is subtracted the two are .07 and $-.07$. It is the policy count that the bar follows, for the reason (6) shows.

What growing a board does, and does not, do. Equation (6) is free of T in its leading term at fixed $\mu(\varepsilon)$ and σ^2, and Fig. 3 draws more tasks from a fixed per-cell law and finds the bar rising and then flattening. Neither says that appending an arbitrary task raises it, and appending one can lower it. Take a task on which every policy scores the board's own grand mean. It leaves S untouched and raises the budget from $T\varepsilon$ to $(T+1)\varepsilon$, so the feasible set of (3) shrinks, d_ε cannot fall and the bar cannot rise. What is monotone is the bar at a fixed total budget, and the bar under replication.

Proposition 6. *Write $d_B = \min\{d(a) : \sum_t \ell_t(a) > B\}$ and $r_B = \sqrt{1 - d_B/S}$ for a fixed total budget $B \geq 0$. If p' appends any column to p then $d_B(p') \leq d_B(p)$ and $S(p') \geq S(p)$, so $r_B(p') \geq r_B(p)$.*

Every choice on p extends to p' by taking a leader on the new task, which leaves its total loss unchanged and costs nothing, the new column lying in its leader's cone, so d_B cannot rise. The enlarged table's own mean minimises its squared deviations, so S cannot fall. The per-task normalisation is therefore the whole of the non-monotonicity in T: at a fixed total budget, adding a task never makes a board easier to certify.

Proposition 7. *Let $p^{(m)}$ repeat every column of p m times. Then $S(p^{(m)}) = mS(p)$, and $d_\varepsilon(p^{(m)}) \leq m\, d_\varepsilon(p)$ when $d_\varepsilon(p)$ is finite, so $r_\varepsilon(p^{(m)}) \geq r_\varepsilon(p)$ in every case.*

Centring is unchanged by replication, so S scales by m; replicating a minimiser of (3) leaves its mean regret and multiplies its cone cost by m, so $d_\varepsilon(p^{(m)}) \leq m\, d_\varepsilon(p)$ and the ratio d_ε/S cannot rise. Replication leaves $\frac{1}{T}\sum_t L_t$ unchanged, so d_ε is infinite on $p^{(m)}$ exactly when it is on p, and both bars are then zero. The inequality can be strict, a replicated board admitting choices that concede on some copies of a task and not others. Replicating a benchmark therefore never lowers the correlation an evaluator must clear. The bar tends to rise with the policy count, for the reason (6) shows: $S = KT\sigma^2$ grows with K while the margins of a fixed per-cell law do not. That direction is not monotone either. On the board with rows $(0,.9)$ and $(.2,.9)$, a third policy at $(.5,1)$ lies above every entry and raises both margins, from .2 to .3 and from 0 to .1; the bar at $\varepsilon = .05$ falls from .985 to .974. Propositions 6 and 7 and the mean-padded column are verified exactly on 1500 grown and replicated cases by `sims/bracket_stress.py`, which checks this witness as it stands; the mean-padded column lowers the bar on 413 of those cases.

Proposition 2 sends the runner-up endpoint, a necessary one, to a limit and the end of its proof sends the envelope endpoint, a sufficient one, but a quantity trapped between two limits need not have one. That the two limits differ is not proved here; it is what the numbers show, the runner-up program being exact exactly when some minimiser of (3) concedes only runner-ups, which a board satisfies less and less often as it lengthens. On the random boards verified below that criterion holds in 1605 of the 2139 comparisons with at most eight tasks and in 3361 of the 6170 with more than twenty. The bar converges all the same, and the proposition below proves it directly rather than through either endpoint.

Proposition 8. *Let the columns of p be independent draws from a fixed law on $[0,1]^K$ whose top-two margin g has a density, and let $0 < \varepsilon < \mathbb{E}[g]$. Then d_ε/S and r_ε converge almost surely as $T \to \infty$, to values fixed by the column law and ε.*

Each column carries the finite rung set $\{(0,0)\} \cup \{(\ell_t(i), d_t(i))\}_i$, and d_ε is a cheapest cover of the budget by one rung per task. Let $\hat{\varphi}_t$ be the convex envelope of that set on $[0, L_t]$, let $\psi_t(\lambda) = \min_{0 \leq \ell \leq L_t}(\hat{\varphi}_t(\ell) - \lambda\ell)$, and let $\Phi_T = \min\{\sum_t \hat{\varphi}_t(\ell_t) : 0 \leq \ell_t \leq L_t,\ \sum_t \ell_t \geq T\varepsilon\}$. Every choice is feasible for Φ_T at no greater cost, so $\Phi_T \leq d_\varepsilon$. The programme is convex and separable with one constraint, and $\ell = L$ is a Slater point once $\sum_t L_t > T\varepsilon$, so strong duality holds; let $\lambda^\star$ be the largest optimal price, and write $\underline{\ell}_t$ and $\overline{\ell}_t$ for the smallest and largest minimiser at it, so that $\Phi_T = \sum_t \psi_t(\lambda^\star) + \lambda^\star T\varepsilon$. That price is at most $\frac{1}{2}$. A task with $g_t > 0$ has first hull edge of slope exactly $g_t/2 \leq \frac{1}{2}$, since a rung of loss ℓ costs at least $\ell^2/2$ by Lem. 1 and the runner-up attains $g_t^2/2$. Above that price its largest minimiser is therefore a nonzero vertex, of loss at least g_t; a task with $g_t = 0$ meets $\overline{\ell}_t \geq 0 = g_t$ regardless. Hence $\sum_t \overline{\ell}_t(\lambda) \geq \sum_t g_t > T\varepsilon$ for $\lambda > \frac{1}{2}$ once the margin mass clears the budget, which $\varepsilon < \mathbb{E}[g]$ makes eventual and almost sure, and the dual is strictly decreasing there.

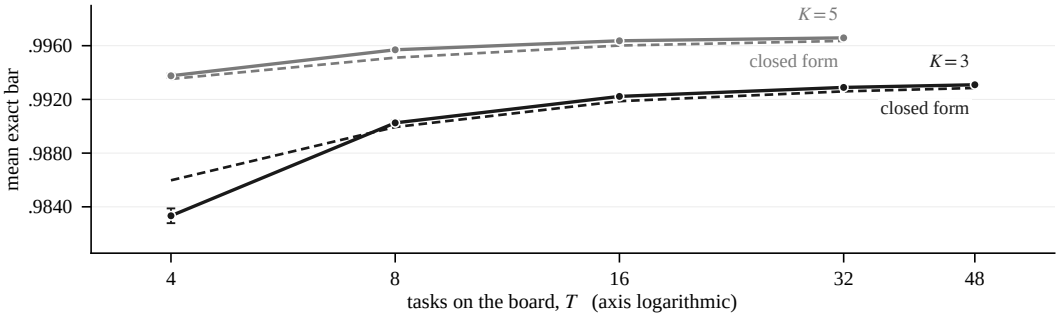


Figure 3: Mean exact bar at $\varepsilon = .05$ as the board grows, over four hundred nested draws of independent uniform cells. The smaller boards are the first columns of the larger, so each step is a paired difference on the same boards, and a step inside its own error would carry no direction whichever way it pointed. Every step is a rise, the smallest by at least 4.3 standard errors, the one-sided lower end of the noncentrality interval taken simultaneously over the seven steps, and each rises on at least .61 of the boards it is measured on. Bars are one standard error of the mean. The closed form of Prop. 2 is the dashed companion of each series, averaged over the draws on which it exists. The axis is logarithmic in the task count, so a rise whose steps shrink by a constant factor is a line that flattens. The five-policy series stops at $T = 32$, the largest of that shape the scan draws.

Each $\hat{\varphi}_t(\ell) - \lambda\ell$ is minimised at a vertex of the envelope, where $\hat{\varphi}_t$ equals the true rung cost. At the largest optimal price $\sum_t \underline{\ell}_t(\lambda^\star) \le T\varepsilon < \sum_t \overline{\ell}_t(\lambda^\star)$, the second inequality strict because equality would leave the dual flat above $\lambda^\star$ and contradict maximality. Moving tasks one at a time from $\underline{\ell}_t$ to $\overline{\ell}_t$ therefore passes $T\varepsilon$ strictly, so some vertex assignment ℓ° has $T\varepsilon < \sum_t \ell_t^\circ \le T\varepsilon + \max_t L_t$. It is a genuine choice, and $d_\varepsilon \le \sum_t \hat{\varphi}_t(\ell_t^\circ) = \sum_t \psi_t(\lambda^\star) + \lambda^\star \sum_t \ell_t^\circ \le \Phi_T + \lambda^\star \max_t L_t \le \Phi_T + \frac{1}{2}$. The gap is bounded independently of T, which is the whole of the argument.

Finally $\Phi_T/T = \sup_{0\le\lambda\le 1/2}\big(\frac{1}{T}\sum_t \psi_t(\lambda) + \lambda\varepsilon\big)$. Each ψ_t is determined by its own column, takes values in $[-\frac{1}{2}, 0]$, and is 1-Lipschitz in λ because its derivative is $-\overline{\ell}_t(\lambda) \in [-1, 0]$. The strong law at finitely many prices and that Lipschitz bound give uniform convergence of $\frac{1}{T}\sum_t \psi_t$ on $[0, \frac{1}{2}]$ almost surely, so Φ_T/T, and with it d_ε/T, converges to $\sup_\lambda(\mathbb{E}[\psi(\lambda)] + \lambda\varepsilon)$. The same law sends $\sigma^2 = S/(KT)$ to the variance of a cell drawn by sampling a column and then a row, positive because g has a density, so d_ε/S and r_ε converge. The released check verifies, on 1090 board-tolerance cases whose margin mass clears the budget, the two finite-sample facts the argument turns on: that the relaxation never exceeds the exact distance, and that neither the price nor the gap exceeds its bound.

Numerical verification. All arithmetic in these checks is exact over the rationals. The random population is 3400 boards with $2 \le K \le 8$, $2 \le T \le 40$ and entries on a grid of $1/10$, $1/100$ or $1/1000$, each priced at $\varepsilon \in \{.01, .02, .05, .10, .20\}$ with d_ε computed exactly. Lemma 1's half square, runner-up identity and midpoint criterion, the necessary side and its equality at three policies, the runner-up program and the sufficient side in its envelope form were checked wherever each applies, with no violation. The necessary side exists on each of the 15483 board-tolerance cases with a costly choice and on no other, and is exact on 8274 of them, among them all 4170 with at most three policies. The worst slack was $(D^+ - d_\varepsilon)/S = .5245$ on the necessary side, .2802 for the runner-up program alone, and $(d_\varepsilon - D^-)/S = .8281$ on the sufficient one. The floor h_t itself is checked against d_ε on every board of Table 2, $D^{\mathrm{disc}}(\varepsilon) \le d_\varepsilon$ being asserted there when the artifact is built.

A.4 Summaries with and without a board, and the null

Corollary 2 (what a correlation gives with no board)**.** *For every $K \ge 2$, every T and every board $p \in [0,1]^{K\times T}$, $\varepsilon^\star(r) \le \sqrt{K(1-r^2)/2}$. The bound needs the policy count and the correlation alone, and is vacuous at or below $r = \sqrt{1 - 2/K}$.*

Let a satisfy $d(a) \le S(1-r^2)$. Lemma 1 gives $d(a) \ge \frac{1}{2}\sum_t \ell_t(a)^2$, and the inequality of Cauchy and Schwarz gives $\sum_t \ell_t(a)^2 \ge \big(\sum_t \ell_t(a)\big)^2/T = T\,R(a)^2$, so $\frac{T}{2}R(a)^2 \le S(1-r^2)$. And S is the total squared deviation of KT numbers in $[0,1]$ from their own mean, so $S \le KT/4$, with equality when

Table 3: Three summaries of the same evidence, priced for the per-task decision. Each certificate other than the closed form is the exact worst regret compatible with publishing that summary and nothing else. All three summaries are recomputed here from the printed columns, the rank statistic under the tie rule (7) states. *Strict* is the certificate of Prop. 3(ii), which the published score table licenses when its taskwise argmax is unique and which is n/a where it is not; *any* is the unconditional certificate of Prop. 3(i), which holds whatever the ties. Both are exact knapsacks, and the applicable one beats the correlation on every pair; *closed* is the factor $\frac{K}{2}v$ or Kv of the same proposition, which needs no board and is never tighter. It loses to the correlation on two pairs: at .301 against .189 on AgentRewardBench's rule-based checker, and by under a thousandth on SIMPLER's variant aggregation. The last column is what the evaluator loses, with ties resolved against it. Values are rounded from their exact rationals, halves upward.

	correlation	rank violations					agreement count		
evaluator and board	$\varepsilon^\star(r)$	MMRV	unique	strict	any	closed	m/T	$\varepsilon^\star(m)$	realised
WorldGym	.447	.0480	yes	.071	.144	.072	16/17	.059	.012
SIMPLER Google, visual match	.335	.0494	yes	.128	.185	.148	3/4	.218	.017
SIMPLER Google, variant agg.	.498	.1661	yes	.310	.380	.498	1/4	.593	.197
SIMPLER WidowX	.165	.0278	yes	.042	.081	.042	3/4	.108	.042
Ctrl-World, instruction	.114	.0000	yes	.000	.000	.000	7/7	.000	.000
Ctrl-World, success	.293	.0190	no	n/a	.057	.057	6/7	.107	.007
AutoEval, autonomous	.096	.0147	yes	.044	.064	.044	5/5	.000	.000
AutoEval, SIMPLER	.870	.2067	yes	.525	.685	.620	1/4	.680	.415
AgentRewardBench, judge	.146	.0286	yes	.044	.091	.057	4/4	.000	.000
AgentRewardBench, rule	.189	.0751	no	n/a	.164	.301	1/4	.166	.071
WMT23 chrF	.132	.0196	yes	.042	.049	.088	1/8	.128	.011

KT is even and half the cells sit at each end. Combining, $R(a)^2 \le \frac{K}{2}(1-r^2)$. The bound is at or above the trivial one exactly when $K(1-r^2) \ge 2$. It is crude (it discards the margins, the ranges and T), but it is what a reader of a benchmark that prints no board is left with, and Sec. 4 prices it there against the rank statistic's. It is not crude enough to explain that comparison away. On T identical columns $(1, 1-\ell, 0, \dots, 0)$ the choice that takes the runner-up everywhere has regret ℓ and, by Lem. 1's runner-up identity, cone cost exactly $T\ell^2/2$, so $\varepsilon^\star(r) \ge \ell$ at the correlation that cost implies. Writing $S_{\mathrm{col}} = 1 + (1-\ell)^2 - (2-\ell)^2/K$, the witness reaches ℓ against a bound of $\ell\sqrt{K/(4S_{\mathrm{col}})}$, a ratio $2\sqrt{S_{\mathrm{col}}/K}$ free of T and least at $\ell = (K-2)/(K-1)$, where $S_{\mathrm{col}} = (K-2)/(K-1)$. Each ℓ pins one correlation, and $\ell \le 1$ caps how low: no member of the family sits below $r = \sqrt{1-K/(2(K-1))}$, which is at or below $\sqrt{1-2/K}$ because $(K-2)^2 \ge 0$. So for every $K \ge 3$ and every r at which the corollary is not already vacuous, the bound is attainable within $\big(4(K-2)/(K(K-1))\big)^{-1/2}$, which at $K=5$ is $\sqrt{5/3}$. Below $r = 1/\sqrt{3}$ no witness could do better at any $K \ge 3$, the bound exceeding the unit range there.

Publishing both summaries. A benchmark that prints a correlation and a rank statistic confines the evaluator to the intersection of the two classes, and a choice is admitted only if one score table meets both. The two per-choice minima sit at different tables (the cone distance at $\Pi_{\mathscr{C}_a}(p)$, the rank charge at a table tying the selected policy with everything above it), so intersecting the two budgets separately relaxes the pair's own exchange rate rather than computing it. Under the strict charge that relaxation is exact. Lifting $\Pi_{\mathscr{C}_a}(p)$ so the selected policy is first on every task carries cone cost arbitrarily close to $d(a)$ and charge exactly c^+, so a choice inside both budgets is realised by one table in the limit, as in Prop. 1. Both budgets are additive over tasks, so the program is the shortest path of (2) carrying two coordinates instead of one, with the Pareto front of the two kept at each attainable loss. It returns exactly the rank statistic's own certificate on all eleven pairs, and on the nine whose published score table has a unique taskwise argmax that value is the pair's own exchange rate and not merely above it. On those nine the correlation adds nothing to what publishing the statistic alone licenses. That is a statement about these boards and not a bound; a board on which the two summaries cut in different directions would admit less than the smaller of the two. This analysis is not registered.

A board-plus-noise null. A margin-weighted rank statistic is well placed to certify a margin-weighted rank quantity, and a count of led tasks to certify a regret carried by the tasks that are not led. So the comparisons of Sec. 4 are worth pricing before any is read as evidence about a benchmark. This matched null holds each pair's board fixed and replaces its evaluator by the board plus Gaussian

noise, clipped to $[0,1]$ and rounded to six places. Its scale is set per pair so that the median correlation a draw induces is that pair's recomputed one, by forty bisection steps on the median of sixty probe draws. Each of five independent streams sets its own scale this way, from probes of its own, and then draws two hundred evaluators per pair. Over all of them the median correlation those draws realise departs from the target by at most .0544, the largest gap over the streams run rather than the gap on one. Every quantity is computed as for the published evaluator. It all but reproduces both results, one pair's factor aside. The rank statistic is the tighter of the two on at least .9430 of draws, the mean over the eleven pairs of the one-sided lower ends of their binomial intervals, taken simultaneously over the eleven. Over the ten pairs with a positive certificate, where alone the factor is defined, the median of the per-pair median factors runs from 2.643 to 2.859 over the streams, against the corpus's 2.884. The corpus exceeds its own pair's null median on six or seven of those ten, according to the stream. The null matches only the median correlation over its calibration draws. It does not match the tie structure: on the two pairs whose published evaluator ties it earns the halved certificate that evaluator is denied, which moves its ratios up rather than down. A random-board null draws the boards at random as well, over six noise scales whose median correlations run from .450 to .995 and so nearly cover the corpus's own. The rank statistic is the tighter of the two on at least .938 of the draws each scale prices, the one-sided lower end of the binomial interval taken simultaneously over the six. The widest scale prices 9954 of its 10000 draws, the rest admitting no correlation certificate. The advantage is not a property of these nine boards either. The agreement count of Prop. 4 is priced against the same null. A noise evaluator's count is the tighter of the two on at least .7538 of draws, read the same way, so its ten of eleven on the corpus is no more evidence about these benchmarks than the rank statistic's eleven. The null is not a claim that these evaluators are noise. It is the statement that, one pair's factor aside, neither comparison distinguishes them from it once the rule-based checker's null draws are priced by the certificate the checker itself gets, and that the conversion, not the corpus, is what separates the two summaries. The third comparison is a count, and it is no better off. Resolving the decision means a certificate at or below the best fixed policy's own regret, which the rank statistic does on three pairs and the correlation on one. Priced against the same null at 4000 draws a pair on each stream, the rank statistic resolves at least 2.273 pairs of eleven in expectation and the correlation at least 1.000. Every pair's probability is read at the lower end of its own interval, and the eleven are unequal. So their sum is Poisson binomial. Its tail rises with every one of the eleven, so evaluating it at their lower ends bounds it from below: three or more is what happens at least .379 of the time. The correlation's one is worse than an empty finding. Every noise evaluator drawn against AutoEval's autonomous proxy resolves that board, the one pair the published correlation resolves, and 29 of 20000 resolve SIMPLER's WidowX arm, so the null's expectation exceeds the corpus's one. So the count does not separate the two summaries either.

Each of the ten pairs with a factor is also tested against its own null. With k of n null draws reaching a pair's factor, its one-sided Monte Carlo p-value is $(k+1)/(n+1)$. WMT23's null reaches that pair's factor once in the first 200 draws of the first stream, and on 36 of all 20000 at 4000 draws a stream: a p-value of .0019, and .019 after Bonferroni over the ten. The next smallest after the correction is .77, on AutoEval's SIMPLER bridge. Ctrl-World's instruction pair is not tested, its rank statistic certifying zero and leaving no factor. In the other tail only AgentRewardBench's rule-based checker lies below its null: 16 of its 19855 draws are at or under its factor, a p-value of .009 after Bonferroni. The checker ties on one task, so its rank certificate is the unconditional one of Prop. 3(i), while a drawn evaluator that does not tie is priced by the strict one. Priced by the unconditional certificate as well, 1617 of those draws are at or under the factor, a p-value of .82 after Bonferroni. Fisher's method combines the ten p-values into one. Pooled over the streams it is .059 at 200 draws a stream and .052 at 4000, and .39 and .42 without WMT23. The combination rests on the one pair the per-pair test already names, so it is not reported as a finding. This analysis is not registered.

A.5 Coarser decisions, and the pooled and within-task conventions

Split a score table orthogonally as $y_{it} = u_i + w_{it}$ with $\sum_t w_{it} = 0$, and the board as $p_{it} = m_i + q_{it}$ likewise. The cross terms vanish, so $\|y-p\|_F^2 = T\|u-m\|^2 + \sum_{i,t}(w_{it}-q_{it})^2$, and any cone constraining u alone has squared distance T times the squared distance from m to its own set in $\mathbb{R}^K$. The aggregate cone $\{y : \bar{y}_i \geq \bar{y}_j \ \forall j\}$ is one, and its distance comes from the pooling of Lemma 1. The shortlist cone $\{y : \min_{i\in A} \bar{y}_i \geq \max_{j\notin A} \bar{y}_j\}$ for a set A of k policies is one too, cut out by $k(K-k)$ inequalities between two means. Both are closed, convex and contain every column-constant matrix, which is all Prop. 1 uses. The shortlist set is the union over a threshold τ of the boxes $\{u_i \geq \tau \text{ on } A,$

$u_j \le \tau$ off it}, so its distance is $\min_\tau \big[\sum_{i\in A}((\tau-m_i)_+)^2+\sum_{j\notin A}((m_j-\tau)_+)^2\big]$, convex and piecewise quadratic in τ with kinks at the coordinates, hence exact over the rationals. The regret of keeping A is $\max_j \bar p_j - \max_{i\in A}\bar p_i$, maximised over the k-subsets the correlation admits; at $k=1$ the two cones coincide and the released check asserts the two values agree on every pair. Table 1 gives it at $k=2$, with the smallest k at which it meets the tolerance. This analysis is not registered.

Proposition 9 (a longer shortlist never costs more)**.** *For every board, every r and every* $1 \le k < K$, $\varepsilon^\star_{k+1}(r) \le \varepsilon^\star_k(r)$.

Let A' be a feasible set of $k+1$ policies and let $j \in A'$ have the least mean in it. Write $A = A' \setminus \{j\}$; the two have the same best member, so the same regret. Let τ attain the distance for A'. If $\tau < m_j$ then every kept term vanishes, and raising τ to m_j leaves them vanishing while no dropped term grows, so $\tau \ge m_j$ may be assumed. There j's own term $((\tau-m_j)_+)^2$ is replaced, when j moves off the set, by $((m_j-\tau)_+)^2 = 0$, and every other term is unchanged. Hence $d(A) \le d(A')$, A is feasible too, and the maximum over k-sets is at least the regret of A'. The containment argument the body might suggest is not available: $\mathscr{C}_A \not\subseteq \mathscr{C}_{A\cup\{j\}}$ in general.

Corollary 3 (the aggregate bar, exactly)**.** *Let m be the vector of policy means,* $\sigma^2 = S/(KT)$*, and* $d_\star(m,i)$ *the squared distance from m to* $\{u \in \mathbb{R}^K : u_i \ge u_j\ \forall j\}$*. The correlation that rules out an aggregate regret above* ε *is*

$$r^{\mathrm{agg}}_\varepsilon = \sqrt{1 - \tfrac{T}{S}\min\big\{d_\star(m,i) :\ \max_j m_j - m_i > \varepsilon\big\}},$$

and if $m_{(1)} - m_{(2)} > \varepsilon$ *the minimum is attained at the runner-up, so*

$$r^{\mathrm{agg}}_\varepsilon = \sqrt{1 - \frac{(m_{(1)}-m_{(2)})^2}{2K\sigma^2}}.$$

Where no policy is costly the family is empty and $r^{\mathrm{agg}}_\varepsilon = 0$*, as in* (3)*.*

The first display is (3) with the aggregate cone and the split above. For the second, Lemma 1 applied to the single vector m gives $d_\star(m,i) \ge (m_{(1)}-m_i)^2/2$ with equality at the runner-up, and every costly policy has $m_{(1)} - m_i \ge m_{(1)} - m_{(2)}$. So where the finer decision needs the whole margin profile and yields a bracket, the coarser one needs the mean row, the task count and S, and yields the bar itself. The same split prices a shortlist of k, the minimum then running over the k-subsets whose regret exceeds ε; by Prop. 9 that bar is nonincreasing in k, and a correlation clears it exactly when the certificate at that k is within the tolerance. At $k=K$ no subset is costly and the bar is zero, so every evaluator of positive correlation clears some k on any board whatever; the question is what it costs, and on AgentRewardBench's rule-based checker it is the whole field of four. Neither this nor the certified shortlist size of Sec. 2 is registered.

Proposition 10 (the within-task convention, priced)**.** *Let* $\rho(y,p) = \frac{1}{T}\sum_t \mathrm{corr}(y_{\cdot t}, p_{\cdot t})$ *be the coefficient computed inside each task and averaged, over tables with no constant column, every column of p being nonconstant as well, and let* $M_t(i) = \max\{\mathrm{corr}(y_{\cdot t}, p_{\cdot t}) : y_{\cdot t} \in \mathscr{C}^t_i$ *nonconstant*$\}$ *be the maximum over one column's cone. For any choice a,*

$$\max\{\rho(y,p) : y \in \mathscr{C}_a\} = \tfrac{1}{T}\textstyle\sum_t M_t(a(t)), \qquad M_t(i) = \sqrt{1 - d_t(i)/SS_t} \ \text{ whenever } d_t(i) < SS_t,$$

the per-column maxima being attained separately and simultaneously. Hence the exchange rate at a published v, $\varepsilon^\star_\rho(v) = \max\{R(a) : \frac{1}{T}\sum_t M_t(a(t)) \ge v\}$*, is an exact knapsack over the same per-task rungs.*

The cone is the product of the per-column cones and ρ is a sum of terms each depending on one column, so the maximum is the sum of the per-column maxima and they are attained at once. Each $M_t(i)$ is a maximum of a linear functional over a compact set, the column cone met with the unit sphere in $\mathbf{1}^\perp$, so it exists. Proposition 1 applied to the single column t, with S there equal to SS_t, evaluates it as the square root whenever $d_t(i) < SS_t$, at $\Pi_{\mathscr{C}_a}(p)$. Where $d_t(i) = SS_t$ that projection is the column's own mean, which no admissible table carries, and only $M_t(i) \le 0$ survives. Both ρ and R are additive over tasks, so the maximum is a two-objective knapsack solved by the dynamic programme of (2) carrying the attainable ρ instead of the cone cost. Square roots are irrational in general, so the released implementation brackets each between two rationals at most 10^{-40} apart. Its

upper pass takes the upper brackets and reads a saturated column's $M_t(i)$ as 0. Both over-state ρ, so that pass admits every feasible choice. Its lower pass takes the lower brackets and reads a saturated column as -1, so it admits only what is feasible whatever those maxima are. The two passes return the same attainable regret on every pair here, which pins the exact value, and the released check asserts it. Reading the pooled curve at the within-task numeral prices neither summary, the two classes of evaluator being incomparable. They can also be far apart as numbers: on AgentRewardBench's rule-based checker the same two printed columns give .9314 pooled and .1576 within each task and averaged. A pooled coefficient is paid for ranking the tasks as well as the policies within them, and only the second is the decision; Corollary 4 prices the first. This analysis is not registered.

Corollary 4 (when a pooled correlation certifies nothing)**.** *For every choice $d(a) \le \sum_t SS_t$, so $\varepsilon^\star(r) = \frac{1}{T}\sum_t L_t$ whenever*

$$r^2 \;\le\; 1 - \tfrac{1}{S}\textstyle\sum_t SS_t \;=\; \tfrac{K}{S}\sum_t (m_t - \bar{p})^2,$$

the share of S that lies between the task means.

The column-mean vector lies in every per-column cone, so $d_t(i) \le SS_t$ and $d(a) \le \sum_t SS_t$ for every choice; where $S(1-r^2)$ reaches that sum the constraint in (2) binds on nothing and the maximum is over all choices, which is the trivial bound. The threshold needs one number beyond S. It is a sufficient condition and not a necessary one: a correlation above it can still certify the trivial bound, as AutoEval's SIMPLER bridge does at $r = .450$. On the corpus the between-task share runs from .1487 to .8683. The one board it explains is AgentRewardBench's, the highest, whose threshold of .9318 the rule-based checker's .9314 sits just below, while the same board's judge at .9648 clears it. A benchmark can check the condition before publishing, from the same $2T+1$ numbers Cor. 1 needs and one more.

A.6 The class restricted to the evaluator's own errors

Equation (2) maximises over every score matrix at the recomputed correlation, a class far larger than any plausible simulator. The answer is not to argue that the maximiser is plausible but to shrink the class to one the published evaluator generates and recompute. Permute, within each task, which policy carries which of that task's errors. Every task keeps its own multiset of error magnitudes exactly, so each member is as accurate per task as the evaluator that was published; what changes is only the attribution. The class is finite and the maximisation over it is exact.

Proposition 11 (the exchange rate restricted to the evaluator's own residual permutations)**.** *Let p be nonconstant, let $e = y - p$, and for $\pi = (\pi_1, \dots, \pi_T)$ with each π_t a permutation of the K policies write $y^\pi_{it} = p_{it} + e_{\pi_t(i)t}$. The class $\mathscr{Y} = \{y^\pi\}$ has one member for each of the $(K!)^T$ profiles π, a table recurring where a task's residuals tie, and contains y at the identity. Put $N = KT$, $A = \langle p, p\rangle$, $E = \langle e, e\rangle$, $\Sigma_p = \sum_{i,t} p_{it}$, $\Sigma_e = \sum_{i,t} e_{it}$, $\Sigma_y = \Sigma_p + \Sigma_e$, $a = A - \Sigma_p\Sigma_y/N$, $b = A + E - \Sigma_y^2/N$ and*

$$C(\pi) \;=\; \textstyle\sum_t \big\langle p_{\cdot t},\, \pi_t(e_{\cdot t})\big\rangle \;=\; \sum_t \sum_i p_{it}\, e_{\pi_t(i)t}.$$

Then, with $\varepsilon^\star_{\mathscr{Y}}(r) = \max\{R(a(y^\pi)) : \mathrm{corr}(y^\pi, p) \ge r\}$:

- *(i) every member has the same cell sum Σ_y and the same residual sum of squares E, and $\mathrm{corr}(y^\pi, p) = \rho(C(\pi))$ for $\rho(C) = (C+a)\big/\sqrt{S(2C+b)}$;*
- *(ii) ρ is quasiconvex on $\{C : 2C + b > 0\}$, falling below $C = a - b$ and rising above it, so $\{C : \rho(C) < r\}$ is a single interval;*
- *(iii) C and $T\,R(a(y^\pi)) = \sum_t \ell_t(\pi_t)$ are both sums of terms in which π enters only through π_t;*
- *(iv) if no member of $\mathscr{Y}$ is constant, then a total loss L is met by some member at the correlation if and only if $\rho(\overline{C}(L)) \ge r$ or $\rho(\underline{C}(L)) \ge r$, where $\overline{C}(L)$ and $\underline{C}(L)$ are the largest and the smallest $C(\pi)$ among the members of total loss L; both are returned by a dynamic program over L carrying two integers per state, and $\varepsilon^\star_{\mathscr{Y}}(r) = \max\{L/T : L \text{ admissible}\}$, attained.*

Proof of Proposition 11. (i) A permutation inside column t permutes that column's residual entries, so $\sum_i e_{\pi_t(i)t} = \sum_i e_{it}$ and $\sum_i e^2_{\pi_t(i)t} = \sum_i e^2_{it}$; summing over t gives $\sum_{i,t} y^\pi_{it} = \Sigma_p + \Sigma_e = \Sigma_y$ and $\langle e^\pi, e^\pi\rangle = E$ for every π. Expanding $y^\pi = p + e^\pi$, $\langle y^\pi, p\rangle = A + C(\pi)$ and $\langle y^\pi, y^\pi\rangle = A + 2C(\pi) + E$. Centring,

$\langle \tilde{y}^\pi, \tilde{p} \rangle = A + C - \Sigma_y \Sigma_p / N = C + a$ and $\|\tilde{y}^\pi\|_F^2 = A + E + 2C - \Sigma_y^2 / N = 2C + b$, while $\|\tilde{p}\|_F^2 = S > 0$; the ratio is $\rho(C)$. So the whole class collapses onto one scalar, and the correlation constraint is a condition on that scalar alone.

(ii) On $\{2C + b > 0\}$ the function ρ is differentiable and $\rho'(C) = \big((2C+b) - (C+a)\big) \big/ \big(\sqrt{S}(2C + b)^{3/2}\big)$, whose numerator $C + b - a$ is increasing in C and vanishes at $C = a - b$. Hence ρ is nonincreasing to the left of $a - b$ and nondecreasing to its right, so it is quasiconvex on that half line and each strict sublevel set $\{\rho < r\}$ is an interval. Equivalently, clearing the square root, $\rho(C) \geq r$ with $r > 0$ says $C + a \geq 0$ and $C^2 + 2(a - r^2 S)C + (a^2 - r^2 S b) \geq 0$, an upward parabola in C whose failure set is one open interval. The derivative's numerator is $\|\tilde{y}^\pi\|_F^2 - \langle \tilde{y}^\pi, \tilde{p} \rangle = \langle \tilde{y}^\pi, \tilde{e}^\pi \rangle$, so the class need not sit on the rising branch, and the falling branch is why both ends are tested in (iv) rather than only the larger.

(iii) The column of y^π at task t is $p_{\cdot t} + \pi_t(e_{\cdot t})$, which depends on π only through π_t, and so therefore do the taskwise argmax of that column and its loss ℓ_t; the cross term is a sum of such column terms by definition.

(iv) Necessity is immediate, the two extremes being attained. For sufficiency, suppose neither extreme meets the correlation. No member is constant, so $\|\tilde{y}^\pi\|_F^2 = 2C(\pi) + b > 0$ for every π and the whole attainable range of C lies in the half line of (ii). Both extremes then lie in the single interval $\{\rho < r\}$, which is convex, so every attainable C at that loss, lying in $[\underline{C}(L), \overline{C}(L)]$, lies in it too and fails as well. No member is constant on any pair of this corpus, which the released artifact records. The hypothesis is not needed either: a constant member has $2C + b = C + a = 0$, so $a = b/2$ and the vertex $C = a - b = -b/2$ is the least attainable C, every member having $2C + b \geq 0$; ρ is then nondecreasing over the attainable range and the larger end alone decides.

For the computation, fix a loss profile $(\ell_1, \ldots, \ell_T)$. By (iii) the π_t are chosen independently, so the largest cross term consistent with that profile is $\sum_t \bar{c}_t(\ell_t)$, where $\bar{c}_t(\ell)$ is the largest $\langle p_{\cdot t}, \pi(e_{\cdot t}) \rangle$ over the permutations of column t whose loss is ℓ; maximising over profiles summing to L gives $\overline{C}(L) = \max\{\sum_t \bar{c}_t(\ell_t) : \sum_t \ell_t = L\}$, and likewise $\underline{C}(L)$ with the minima. That is a knapsack over tasks, solved by one pass per task over the reachable totals. Enumerating each column costs $O(K!\,K)$ and yields at most K loss levels, so the whole program is $O\big(T(K!\,K + MK)\big)$ with M the number of reachable totals. Scaling p and y to a common integer grid makes every loss, every cross term and every comparison in (ii) an exact integer or rational, ties among an evaluator's maxima included, which a floating point comparison would break by rounding noise on the coarser boards. Each state carries the assignment attaining its extreme, so the maximiser is returned with the value. □

Which tables the class contains. A member of $\mathscr{Y}$ may carry a cell outside $[0, 1]$. The released value maximises over the class as written; one verdict depends on that. A correlation and the choice a score matrix induces are both invariant under $y \mapsto \alpha y + \beta \mathbf{1}$ with $\alpha > 0$, so the affine map onto $[0, 1]$ carries any member there while moving neither the constraint nor the regret. The maximiser is therefore exhibited as an outcome table, though not in general as a member of the class, since the map changes its errors. Clipping each cell into $[0, 1]$ instead is not affine, breaks the reduction of (i), and defines a different class. That class is priced by enumeration on the six pairs where it is determined exactly, and exceeds the released value on none of them. On six of the eleven pairs the members that never leave $[0, 1]$ already attain the released maximum on their own.

Why the noise class cannot carry this argument. The other class a reader will propose is the board plus Gaussian noise at the pair's own calibrated scale. That restriction is vacuous. Let a attain (2) with $d(a) < S(1 - r^2)$. Proposition 1 makes $\mathrm{corr}(\Pi_{\mathscr{C}_a}(p), p) = \sqrt{1 - d(a)/S} > r$, and a small perturbation gives task $a(t)$ a strict maximum while keeping the correlation above r; the affine map above then places the result in $[0, 1]$ without moving either. That table lies in $U = \{u : \mathrm{corr}(u, p) > r,\ u_{a(t)t} > u_{it}\ \forall i \neq a(t), t\}$, which is open and nonempty, and a Gaussian with positive definite covariance gives every nonempty open set positive mass, so the noise class meets U; every member of U induces a and has regret $R(a)$. The noise class therefore certifies exactly $\varepsilon^\star(r)$, on all eleven pairs of this corpus. A class with the right support and the wrong shape restricts a maximum not at all, which is why the restriction reported here is the permutation class, every member of which reproduces the published evaluator's per-task error magnitudes exactly.

What the restriction leaves. Table 4 gives each pair's restricted maximum, and Sec. 2 the range of the share of $\varepsilon^\star(r)$ it keeps. The two ends of the range are Ctrl-World's instruction criterion, where the class is confined to .0071 against a ceiling of .1143, and SIMPLER's variant aggregation, where it still reaches .3468. The maximisations are exact and deterministic, with no sampling: the largest class here is WMT23's $(9!)^8$ tables, and the program that maximises over it exactly visits 353 states. The smaller extreme of the two ends tested in (iv) never decided a level the larger one did not on this corpus, and both are tested because (ii) does not promise otherwise.

One permutation for every task. A within-task permutation keeps each task's error magnitudes and destroys the per-policy structure of the errors, and a reader may hold that the structure is the whole point: an evaluator that overrates one policy overrates it everywhere. Impose that. Restrict $\mathscr{Y}$ to $\pi_1 = \cdots = \pi_T$, one permutation of the K policies applied to every task. Each member then relabels the policies' error profiles among themselves and changes nothing else, a subclass a reader can name without positing a model of simulator error, and it still contains the published evaluator at the identity. The subclass has $K!$ members, so it is enumerated whole and its maximum is exact with no reduction at all. It keeps between .000 and .542 of $\varepsilon^\star(r)$, median .283, and the conclusion still holds: its maximum exceeds the best fixed policy's regret on eight of the eleven pairs, equals it on two and falls below it on one. The one is AutoEval's autonomous proxy, which is already the exception unrestricted. The two are SIMPLER's WidowX arm, where the subclass reaches its best fixed policy's .042 exactly, and Ctrl-World's instruction criterion, where the correlation admits one of the six permutations, the identity, so that entry prices the published evaluator alone, at its own regret of zero. Conditioned on meeting the correlation, the share of the subclass whose regret exceeds the best fixed policy's runs from .000 to .827, median .067, against the within-task class's .262, and on AgentRewardBench's judge the verdict rests on one admissible member of fifteen. Table 4 prints the maximum. So the part of the claim the single-permutation subclass does not carry is the part the corpus already flags. Kept inside $[0,1]$, though, the subclass's maximum exceeds the best fixed policy's regret on five pairs, equals it on three and falls below it on three: WorldGym and SIMPLER's visual match drop below it, and Ctrl-World's success board falls to it.

How typical the within-task maximum is. A maximum is one member, and the within-task class is finite, so the share of it that behaves like the maximum is a count rather than an impression. Condition on the members meeting the pair's correlation and ask what fraction of those carry a regret above the best fixed policy's. Both TR and C are sums of per task terms by Proposition 11(iii), so the joint law of the pair is the convolution of T finitely supported laws and the share is exact. Splitting the tasks into two blocks, convolving each block whole and counting one against the sorted other gives it on ten of the eleven pairs, and only WMT23's $(9!)^8$ members put the convolution out of reach. That share runs from .000 to .850, median .262; it is above a tenth on seven pairs and above a quarter on six. WMT23's is the one sampled value, at 200000 uniform draws of which 82911 meet the correlation, and its one-sided interval puts it at .847 or above. Two pairs need stating plainly rather than averaging away. On SIMPLER's WidowX arm the verdict rests on eight of the 728 admissible members, a share of .011: the maximum clears the baseline there, and almost nothing else in the class does. On AutoEval's autonomous proxy the share is zero, and exactly rather than nearly: the class maximum is .056 against a best fixed policy's .236, so no member can clear it, which is the same exception read a second way.

The verdict under each range convention. The count of pairs whose restricted maximum exceeds the best fixed policy's regret is a property of the convention and is reported under each of the three. As the class is written it is ten of the eleven. Confined to the members that never leave $[0,1]$ it is nine, and under the clip it is nine as well, determined on all eleven pairs. On the six the enumeration reaches, it is read off directly. Elsewhere the in-range maximum already exceeds the baseline and every in-range member is its own clip. The exception is AutoEval's autonomous proxy, where a clipped member is still a score table and Proposition 1 therefore holds its regret at or below $\varepsilon^\star(r) = .096$, under that board's .236. The pair that moves is SIMPLER's WidowX arm, whose in-range and clipped maxima are both .0418, its best fixed policy's regret exactly, rather than the .0628 the class as written reaches; the released maximiser there carries a cell at $-.083$. No other within-task verdict changes under any of the three.

Table 4: The worst case restricted to the evaluator's own errors, read three ways. $\varepsilon^\star(r)$ maximises over every score matrix at the recomputed correlation; *within task* permutes which policy carries which of a task's errors, task by task; *one permutation* applies a single permutation of the policies to every task, which keeps the evaluator's per-policy error structure as well. The restricted entries are taken over the class as written, whose members may leave $[0,1]$; kept inside it, SIMPLER WidowX's within-task maximum falls to its best fixed policy's regret. Kept inside $[0,1]$, the one-permutation maxima of WorldGym, SIMPLER's visual match and Ctrl-World's success board also fall to or below theirs. *Share* is the fraction of the within-task class, conditional on meeting the correlation, whose regret exceeds that board's best fixed policy. It is exact on ten pairs and estimated from 200000 draws on WMT23. The certificates and the best fixed regret are rounded from their exact rationals with halves upward, the share to the nearest place.

	worst case over			how typical	
evaluator and board	$\varepsilon^\star(r)$	within task	one permutation	share	best fixed
WorldGym	.447	.100	.065	.606	.024
SIMPLER Google, visual match	.335	.171	.148	.042	.097
SIMPLER Google, variant agg.	.498	.347	.234	.262	.097
SIMPLER WidowX	.165	.063	.042	.011	.042
Ctrl-World, instruction	.114	.007	.000	.328	.000
Ctrl-World, success	.293	.043	.043	.452	.007
AutoEval, autonomous	.096	.056	.052	.000	.236
AutoEval, SIMPLER	.870	.515	.430	.040	.255
AgentRewardBench, judge	.146	.053	.041	.160	.009
AgentRewardBench, rule	.189	.103	.071	.609	.009
WMT23 chrF	.132	.041	.023	.850	.006

B THE UNCERTAIN-ANCHOR HINGE

For a board p and a radius $\eta \geq 0$ put $u^-_{it} = \max(0, p_{it} - \eta)$ and $u^+_{it} = \min(1, p_{it} + \eta)$, and let $\mathscr{U}_\eta$ be the box of tables lying entrywise between them. A choice a is *certified over the box* at tolerance ε when its regret is at most ε for *every* board in $\mathscr{U}_\eta$. The worst board raises every rival to its ceiling and lowers the chosen policy to its floor, independently by task, so the largest regret any member of $\mathscr{U}_\eta$ assigns to a is $W_{\mathscr{U}_\eta}(a) = \frac{1}{T}\sum_t \kappa_{a(t)t}$, with $\kappa_{it} = \max(0, \max_{j\neq i} u^+_{jt} - u^-_{it})$ the worst taskwise loss of policy i. The relevant quantity is the least regret the box forces on any choice,

$$M_{\mathscr{U}_\eta} = \frac{1}{T}\sum_{t=1}^{T}\Big(\min_i \max_{j\neq i}\big(u^+_{jt} - u^-_{it}\big)\Big)_+ = \frac{1}{T}\sum_{t=1}^{T}\min_i \kappa_{it}, \tag{9}$$

because no choice is certifiable at tolerance ε once $M_{\mathscr{U}_\eta} > \varepsilon$.

Lemma 2 (the uncertain-anchor hinge)**.** *The minimum in* (9) *is attained at the board's own leader. If $p_{(2)t} + \eta \leq 1$ and $p_{(1)t} \geq \eta$ for every task, then*

$$M_{\mathscr{U}_\eta} \;=\; \frac{1}{T}\sum_{t=1}^{T}\big(2\eta - g_t\big)_+. \tag{10}$$

The map $\eta \mapsto M_{\mathscr{U}_\eta}$ is nondecreasing and piecewise linear, with breakpoints only at $g_t/2$, $1 - p_{(2)t}$ and $p_{(1)t}$, so the largest admissible radius $\eta^\star(\varepsilon)$ is attained, and is rational when the board is.

Proof of Lemma 2. Let $i^\star$ be a leader of column t and let $i \neq i^\star$; then $\{j \neq i\}$ contains $i^\star$, so $\max_{j\neq i} u^+_{jt} = \max_j u^+_{jt} \geq \max_{j\neq i^\star} u^+_{jt}$, while $u^-_{it} \leq u^-_{i^\star t}$ because the clip is nondecreasing, whence $\kappa_{it} \geq \kappa_{i^\star t}$. Substituting, and using that the largest entry other than a leader is $p_{(2)t}$, $\min_i \kappa_{it} = \max(0, \min(1, p_{(2)t} + \eta) - \max(0, p_{(1)t} - \eta))$, which is $(2\eta - g_t)_+$ when neither clip binds. Some choice is certifiable at tolerance ε through the box exactly when $M_{\mathscr{U}_\eta} \leq \varepsilon$, so away from the clips that depends on the board only through the margin vector. As a function of η, each $\min_i \kappa_{it}$ is the positive part of a nondecreasing piecewise linear function with kinks at $\eta = 1 - p_{(2)t}$ and $\eta = p_{(1)t}$, and its zero crossing is at $\eta = g_t/2$. The crossing precedes both kinks, since $g_t \leq p_{(1)t}$ and $g_t \leq 1 - p_{(2)t}$, so the breakpoints are among the three listed values and $M_{\mathscr{U}_\eta}$ rises continuously from $M_{\mathscr{U}_0} = 0$ to $M_{\mathscr{U}_1} = 1$. The largest η with $M_{\mathscr{U}_\eta} \leq \varepsilon$ is therefore attained, and is found by locating the two consecutive breakpoints straddling it and solving one affine equation, which returns a rational number whenever the board is rational. □

Table 5: The largest radius $\eta^\star$ at which some choice is certifiable at a five-point tolerance for every board the box admits, computed exactly by Lem. 2. Beside it are the board's mean top-two margin, the number of tasks whose margin the hinge activates at that radius, and the hinge itself at each declared level of the per-cell binomial box of Sec. 4. Every radius is an exact rational, as the released artifact carries it, printed here to four places with halves rounded upward; the margins and hinges are rounded to the nearest place. The hinge is given for the four designs whose source states a per-cell trial count, and n/a marks the rows whose source states none. The AutoEval SIMPLER board states fifty trials too, but its real cells are four of the AutoEval autonomous board's five tasks, so its row gives that design's hinge on those four tasks alone, not a fifth design's. A cell's success count is its stated trial count times its printed rate, rounded to the nearest integer with halves to the even one, and App. B gives the hinge over every count the rates admit where that rule decides. A board admits a certifiable choice exactly when its hinge is at or below .05, which at the wider level none of the four is. The AutoEval SIMPLER row is within the tolerance at both levels: the one task it omits, cloth folding, is what lifts the autonomous board over the tolerance at the wider level.

board	$\eta^\star$	mean g_t	tasks the hinge activates	hinge at 68%	hinge at 95%
WorldGym	.2028	.465	9/17	.0437	.2090
SIMPLER Google Robot	.0645	.102	3/4	.0366	.1445
SIMPLER WidowX	.0960	.144	3/4	.0651	.2166
Ctrl-World, instruction	.0875	.271	2/7	n/a	n/a
Ctrl-World, success	.1000	.264	2/7	n/a	n/a
AutoEval, autonomous	.0860	.364	3/5	.0236	.0798
AutoEval, SIMPLER	.1133	.445	2/4	.0055	.0481
AgentRewardBench	.0593	.073	2/4	n/a	n/a
WMT23	.0299	.010	8/8	n/a	n/a

Equation (10) is Lem. 1's reduction reached from the other side. Where a clip binds, (10) charges more than the box does, so reading the radius off the margins alone is conservative below.

A radius that varies by cell, and where the leader step survives. Lemma 2 locates the inner minimum at the board's own leader using $u^-_{it} \le u^-_{i^\star t}$ and $\max_{j\neq i} u^+_{jt} = \max_j u^+_{jt}$, both of which need the corners to be nondecreasing in the board entry. A uniform radius gives that; a radius varying by cell need not, and the minimum can sit at a non-leader. The sampling boxes of Sec. 4 are nonetheless safe, because within a column they use one trial count, so each corner is a function of the entry alone and is nondecreasing in it. The released implementation does not rely on the argument in any case: it enumerates the inner minimum. The box is also a product set: every cell moves independently, with no joint constraint from shared trials.

Verification. The released check draws small boards and enumerates every choice and every vertex of each box. For each of 2^{KT} corner tables and each of K^T choices it compares the closed expression against the regret computed by brute force. It then confirms three facts at each radius. The per-task minimum is attained at the board's own leader. The closed expression in each column's top two entries, which reduces to (10) when no clip engages, equals the enumerated capacity. And the robust correlation cutoff equals one exactly when $M_{\mathscr{U}_\eta} > \varepsilon$. The check covers 160 box-and-configuration pairs over 32 boards, and passes on all of them.

The radii, and what orders them. Table 5 reports $\eta^\star(.05)$ for the nine boards beside the mean top-two margin and the number of tasks the hinge activates. The radii the corpus admits at a five-point tolerance run from .0299 on WMT23 to .2028 on WorldGym. WorldGym has the largest of both. But Ctrl-World's instruction board has a larger mean margin than SIMPLER WidowX and a smaller radius, while its success board has a larger mean margin and a larger radius. AutoEval's SIMPLER board has nearly WorldGym's mean margin and just over half its radius. Ordering the boards by the mean of the lowest quartile of margins does not reproduce the ordering either.

What the radius does not say. A radius is a declaration about the board, not a measurement of it. The certificate is conditional on that declaration: it says which choices survive every table within η, not that the true rates lie within η of the printed ones. Section 4 takes the handle most of these sources offer, the trial counts they state, and finds them insufficient to resolve the decision on three of the four designs that state any. A fifth board states one, AutoEval's SIMPLER grid at the same fifty trials per cell, and it resolves its own decision too, at .993; Table 5 says why it is not a fifth design.

The attainable correlation is itself capped. A correlation computed against a board estimated from few trials has its own ceiling depressed by the board's noise (Spearman, 1910; Brown, 1910). Comparing a published r against r_ε therefore compares an attenuated quantity against a bar the attenuation may put out of reach, so part of the shortfall may be the board's noise rather than the evaluator's.

Which box prices the sampling. The retrospectively registered criteria of App. D fixed this box as a plug-in multiple of each cell's standard error rather than the exact interval reported here. The two disagree on SIMPLER WidowX, which the plug-in box at one standard error would certify at a five-point tolerance where the exact interval at the corresponding level does not. The plug-in radius is identically zero wherever a cell reports no successes or no failures, which is why the exact interval is the one Sec. 4 reports; the registered criterion is adjudicated on the registered box and is confirmed there too. Two of that board's twelve printed rates, moreover, reconcile with no success count out of the twenty-four trials its source states, and both reconcile with thirty. Had that one task run at thirty, its hinge here would read .0497 rather than .0651 and the board would admit a certifiable choice, by less than .0003 and only at levels below .6809, and its decision-safe probability would read .938. Its certificates, which take no trial count at all, and every registered criterion stand unchanged either way.

How a count is read from a rate. The box needs each cell's success count, and the released check reads it as the stated trial count times the printed rate, rounded to the nearest integer with halves to the even one. Where some count reproduces the printed rate that is the count, and on SIMPLER's Google Robot and AutoEval's boards every cell's does. On WorldGym thirteen rates are half-trial credits, and on SIMPLER WidowX two rates reconcile with no count out of twenty-four; there both neighbouring integers are readings of the rate. Over all 8192 readings of WorldGym's thirteen, its hinge at the 68% level runs from .0382 to .0653 and certifies on 5232 of them. Over the four readings of WidowX's two it runs from .0486 to .0652 and certifies on one. At the 95% level no reading of either board certifies.

Where the clips bite. Equation (10) holds when no task's second-best entry exceeds $1-\eta$ and no task's leader falls below η. Seven of the nine boards satisfy both at the radius Table 5 reports. AutoEval's two do not: one task on each puts its runner-up within η of one, so the ceiling clips and that task's hinge falls below what its margin alone would charge. The released check tests both conditions on every board and claims the margin identity only where they hold.

C THE CORPUS

The corpus holds eleven (board, evaluator) pairs. They come from nine distinct boards and eight source tables, drawn from four robotics benchmark suites, one web-agent benchmark and one machine-translation campaign, because two boards carry two evaluator columns each and one source table carries two boards. Each source table is a file of the archive, and may transcribe more than one published table. Table 6 lists the pairs' shapes, ties and trial counts. Every rate is the value its source prints, divided by the stated trial count where the source prints successes (WorldGym, AutoEval) and by one hundred where it prints a score on a hundred-point scale (AgentRewardBench, WMT23). Only chrF enters rounded, to four places. The printed decimals fix the denominators, and every certificate is exact over the rationals.

WorldGym. The source rolls policies out inside a learned video world model and compares the result with real Bridge robot outcomes (Quevedo et al., 2026). It reprints those outcomes from OpenVLA's evaluation (Kim et al., 2024). The board is the `real` column of `worldgym_v3_tab5_paired.csv` and the evaluator is the `wm` column. Both cover three policies, labelled `rt1x`, `octo` and `ovla`, on 17 tasks, for 51 cells per column. Each cell is scored from ten real trials, with half-trial partial credit allowed. No entry needs more than two decimal places, and all 51 board entries are multiples of $1/20$; thirteen of them are odd multiples, reachable only through half credit. The least common multiple of the board's denominators is therefore 20. Averaged over tasks, the board gives 18.529%, 20.000% and 70.588% for `rt1x`, `octo` and `ovla`. The evaluator gives 15.588%, 23.824% and 67.353% for the same three. Each of these six means reproduces the corresponding published mean to within .089 points.

SIMPLER Google Robot. The source re-runs real-world manipulation policies in simulation and reports two protocols, visual matching and variant aggregation (Li et al., 2024). The board is the `real` column of `simpler_google_robot.csv`, six policies on four task families, for 24 cells. Two evaluator columns share it: `sim_visual_matching` and `sim_variant_aggregation`. This is one of the two boards in the corpus that support two pairs. All three columns are printed to three decimal places, so each rate is a rational whose denominator divides 1000, and the board's least common multiple is 1000. The file states the real trial plan per task in `n_planned`. It is 75 for picking a coke can and 60 for moving near a target. It is 54 for opening or closing a drawer, and 27 for opening the top drawer to place an apple. Every board rate is expressible as a count over its own task's plan, rounded to three decimals.

SIMPLER WidowX. The same source reports a second, smaller grid on a WidowX arm (Li et al., 2024). The board is the `real` column of `simpler_widowx.csv` and the evaluator is `sim_visual_matching`, three policies on four tasks, for 12 cells. Both columns are printed to three decimal places, so the board's denominators again divide 1000. The source states 24 real trials per cell; the file itself carries no count column.

Ctrl-World. The source rolls policies inside a controllable generative world model and prints its rates beside real-world rates (Guo et al., 2026). Three policies, labelled `pi0`, `pi0-fast` and `pi0.5`, are scored on seven tasks under two separate criteria. This yields two boards from one table: `real_instruction` against `wm_instruction`, and `real_success` against `wm_success`. Each board has 21 cells. All four columns are printed to at most two decimal places and every one of the 84 entries is a multiple of $1/20$, so each board's least common multiple is 20. The table states no trial count for any cell.

AutoEval. The source evaluates real robot policies without a human resetting the scene, and prints the autonomous result beside a human-supervised one on the same hardware (Zhou et al., 2025). Six policies, labelled `MiniVLA`, `Octo`, `Open pi0`, `OpenVLA`, `SuSIE` and `SuSIE-LL`, are scored on five tasks. The board is the `real` column of `autoeval_bridge.csv` and the evaluator is `proxy`, for 30 cells. The same source reports the SIMPLER simulator on the four of those tasks SIMPLER models, dropping cloth folding, which gives `autoeval_simpler.csv` and a second board of 24 cells whose real entries are identical to the first board's on those four tasks. Both files carry the integer success count behind every rate, out of the fifty trials each cell plans. All 54 rates in the two boards and all 54 in the two evaluator columns are therefore checkable against their own numerators, and every one of them checks out. No other file in the corpus carries its success counts.

AgentRewardBench. The source scores web-agent trajectories with automatic evaluators and with expert human annotation, and prints both (Lù et al., 2025). The board is the `real` column of `agentrewardbench_table6.csv`, four agents on four benchmarks, for 16 cells, and two evaluator columns share it: `judge`, a language-model judge, and `rule`, each benchmark's own rule-based checker. It is the other board that supports two pairs. All three columns are printed in per cent to one decimal place and enter divided by one hundred. The board's least common multiple is therefore 1000. The tasks here are whole benchmarks rather than single manipulations, so the per-task decision is the choice of an agent per benchmark, and the source states no per-cell denominator.

What each source publishes. Only one of the eleven correlations is a number its own source prints. WorldGym plots one point per policy and task and reports a Pearson coefficient over that pooled scatter, which is the convention Table 1 uses throughout. SIMPLER computes a Pearson coefficient within each task, across policies, and averages it over tasks, and AutoEval follows SIMPLER; on AutoEval's board that average is the .942 its text reports, against the .992 of the pooled scatter. AutoEval also prints SIMPLER's rank statistic, .015, which the recomputed .0147 of Table 3 reproduces. Ctrl-World prints no coefficient at all, AgentRewardBench reports precision and recall against expert annotation rather than a correlation, and WMT23's findings paper reports none either, that comparison living in the companion metrics task. Where a source does report one, its own task grouping need not be this board's: SIMPLER's headline Google Robot average is over three task columns where this board prints four. The r column is therefore recomputed here under one convention for all eleven pairs, and Sec. 2 prices the other.

Table 6: The eleven (board, evaluator) pairs. *Zero-margin tasks* counts the tasks on which the board's best two policies are exactly equal, so no choice between them costs anything. *Evaluator ties* counts the tasks on which the evaluator's argmax is not unique. Trial counts are those the source states; the SIMPLER Google Robot plan varies by task, in the column order picking a coke can, moving near a target, opening or closing a drawer, and opening the top drawer to place an apple. Every number in the table is an exact count.

evaluator and board	source table	policies	tasks	zero-margin tasks	evaluator ties	real trials per cell
WorldGym	`worldgym_v3_tab5_paired`	3	17	0	0	10
SIMPLER Google, visual match	`simpler_google_robot`	6	4	0	0	75/60/54/27
SIMPLER Google, variant agg.	`simpler_google_robot`	6	4	0	0	75/60/54/27
SIMPLER WidowX	`simpler_widowx`	3	4	0	0	24
Ctrl-World, instruction	`ctrlworld_table3`	3	7	2	0	none stated
Ctrl-World, success	`ctrlworld_table3`	3	7	1	2	none stated
AutoEval, autonomous	`autoeval_bridge`	6	5	0	0	50
AutoEval, SIMPLER	`autoeval_simpler`	6	4	0	0	50
AgentRewardBench, judge	`agentrewardbench_table6`	4	4	1	0	none stated
AgentRewardBench, rule	`agentrewardbench_table6`	4	4	1	1	none stated
WMT23 chrF	`wmt23_general_mt`	9	8	0	0	none stated

What the corpus excludes. RoboArena (Atreya et al., 2025) is not used here, because its comparisons are pairwise and decentralised and it publishes no policy-by-task outcome table of the kind the certificate conditions on. AgentRewardBench reports a fifth benchmark, VisualWebArena, on which one of the four agents is absent; including it would break the rectangle, so it is dropped and the board is the four benchmarks all four agents are scored on. The random boards of Sec. 3 are drawn rather than published, and are no part of the corpus either.

Reservations. Three pairs are not robotic. Two of them are AgentRewardBench's, whose board is an expert-annotated success rate rather than a physical trial outcome, and whose tasks are whole benchmarks. The per-task decision there is the choice of an agent per benchmark, which a practitioner does make; but the tasks are not exchangeable with one another, and the scale is not a robot's. The third is nine systems on eight language pairs of WMT23's general machine-translation task. The board is each system's human score under one protocol, DA+SQM, direct assessment with scalar quality metric guidelines, and the evaluator is chrF (Kocmi et al., 2023; WMT Conference, 2023). Both are printed on a hundred-point scale and enter divided by one hundred, the evaluator rounded to four places. The systems are those present in all eight published sub-tables. Six language pairs for which the source published no human ranking are excluded, as is the human reference, against which chrF would be trivially perfect. The source reports how many ratings each language pair received, but a rating on a hundred-point scale is not a binary trial. So the binomial box of Sec. 4 does not apply, and no sampling statement is made about this board either.

Five properties of these files limit what can be said about them. First, the SIMPLER WidowX board does not reconcile cleanly with 24 trials. Seven of its twelve rates are exact multiples of $1/24$, and three more round to such a multiple at three decimals. The remaining two, .233 and .433, are not reachable from 24 trials at any rounding; both sit on the eggplant task, and both are consistent with 30. The matching evaluator column is worse: four of its twelve rates cannot come from 24 trials, and the smallest count consistent with all four is 72. Second, the SIMPLER Google Robot file carries an `early_stopped` flag and an `n_completed` column that does not everywhere match `n_planned`. Exactly one cell disagrees numerically: the drawer task under the policy `RT-1 (Begin)` plans 54 trials and records 2 completed, a shortfall of 52, and it is the only cell whose flag is set. Exactly one further cell leaves `n_completed` blank, the apple task under the same policy, where all three rates are zero. The stated counts describe the board alone: the two simulator columns fail to be counts over the real plan in 17 and 20 of their 24 cells. Depths are the trial counts each source plans, which for one grid exceed the counts it reports completing; the standard errors here are therefore the smallest its design allows. The read is simulated from the printed rates as if they were the truth and the trials independent, and on WorldGym as if its half-trial credits came from whole trials; it prices the sampling the design admits, not the dependence a real evaluation campaign carries. Campaigns share operators, scenes and days, so these are the most favourable numbers the designs permit. The Google Robot board also merges the three orientations of the coke can and the two drawer sub-tasks its source scores separately, so a per-task statistic computed here averages what that source reports apart. Third, the Ctrl-World table states no denominator at all, so no sampling statement can be made about either of its boards. Fourth, AutoEval's two boards are not independent:

Table 7: What two published summaries certify when the board is not published. Thirteen settings from Wang et al. (2026), which prints a correlation and a mean maximum rank violation for each and no outcome table, at five policies throughout. Both certificate columns are bounds from what that source prints and nothing else: *from r* is Cor. 2 capped at the trivial bound, and *from MMRV* the unconditional factor Kv of Prop. 3(i), the only one the source licenses. A certificate at or above 1 is vacuous, regret being a mean of differences of rates in $[0, 1]$. The conversion reads that source's statistic as (7), its own formula save at ties. App. C gives the transcription, the row rule and the one dependent row. The correlations and statistics are printed as the source prints them, the certificates from MMRV exactly, and those from r rounded up.

setting	r	MMRV	from r	from MMRV
VLA-Arena vision	.241	.188	1.000	.940
VLA-Arena layout	.858	.016	.813	.080
VLA-Arena language	.987	.000	.255	.000
VLA-Arena behaviour	.814	.035	.919	.175
SIMPLER vision	.701	.102	1.000	.510
SIMPLER layout	.420	.056	1.000	.280
SIMPLER language	.086	.226	1.000	1.130
REALM vision	.672	.046	1.000	.230
REALM layout	.865	.008	.794	.040
REALM language	.722	.042	1.000	.210
REALM behaviour	.882	.023	.746	.115
REALM before fine-tuning	.785	.030	.980	.150
REALM after fine-tuning	.878	.015	.757	.075

the second is the first restricted to the four tasks its SIMPLER comparison covers, with identical real cells. A statement made about both is therefore one observation and not two, and the head-to-head of Sec. 2 is stated on the four shared tasks for that reason. Fifth, AutoEval's autonomous proxy and its real column are read on the same robot in the same scenes, so the agreement between them is not the agreement of two independent measurements; what the certificate prices is the published pair, whatever produced it. Against that, the structural checks are clean. No board has a duplicate policy name or a duplicate task name, every board is a complete rectangle with no missing cell, and every rate in all eight files lies in $[0, 1]$.

Four of the eight source tables are checked against a transcription independent of the file, not of its transcriber, cell by cell in the order the file prints: WorldGym in its real column alone, and AutoEval's two and AgentRewardBench in every column. The other four are checked for structure only, and no evaluator column of theirs is independently transcribed. On AutoEval's two tables every rate in both columns is also reproduced exactly from the integer count printed beside it. Every corpus quantity reported in this paper is a function of these eight files and the code that reads them, and the reproduction entry point is `code/regenerate.py`.

The board-free population. Table 7 prices a second population, disjoint from the eight source tables above and admitted on the opposite criterion: a source that publishes a correlation and a rank-violation statistic and *no* outcome table. It is not part of the registered corpus, adjudicates no criterion, and was assembled after the criteria had been adjudicated. Wang et al. (2026) is the one source we found that publishes enough of the pair to price. Its Tables 2 and 4 give thirteen settings at five policies, being every row of the two that is not an average over other rows, together with the one average Table 4 reprints and names as a condition of its own. These thirteen pairs are transcribed from their source's own tables and cannot be checked against a board, there being none, and one of the thirteen is that source's own average of four others in the table, so they carry twelve distinct observations. The policy count is the one that source states, and whether its taskwise argmax is unique it does not say, so the weaker of the two factors is the one applied. That source prints a rank correlation beside the two, which is not priced here. A rank correlation constrains the ordering of the cells and not the outcome differences regret is measured in, so any bound from one would need the board it is meant to replace. Two further sources publish the same pair without a table and are left out. Zhang et al. (2026b) do not state how many checkpoints their correlation ranks, which the factor needs, and Gemini Robotics Team (2025) prints its correlation and rank violation inside figure panels rather than a table we could transcribe.

D RELEASED CODE AND OUTPUTS

Layout. The archive holds `MANIFEST.md`, the manuscript under `iclr/` with its sources, figures, compiled PDF and bibliography, and everything else under `code/`. `code/evalceil/` carries the three engines: `evalceil/frontier.py` solves (2) and (3), `evalceil/margins.py` implements the bracket, the summaries' certificates, the accuracy programs of Prop. 5 and the hinge of Lemma 2, and `evalceil/restricted.py` solves the maximisation of Prop. 11. Ten scripts under `code/sims/` write the seven computed artifacts and draw the three figures. The programs in all three engines are exact over the rationals and the integers. Floating point enters where a table or a correlation arrives as a float, which is converted to a rational or refused within a guard of a boundary, and where a result is output. It also enters two computations: a screen whose survivors an exact test then decides, and the hinge of the exact binomial box, whose corners are irrational quantiles.

Registered criteria. `PREREG_margins.md` and `PREREG_anchor.md` fix the population, the tolerance and the kill criteria of each analysis before its outcome is reported, and each script prints its own adjudication. Both registers are retrospective: they were written after these tables had been inspected in earlier work, so what they fix is the adjudication and the population, not a first look at the data. The criteria are P-M1 to P-M6 for the exchange rate, the bracket and the bar, the summaries and the accuracy restriction, P-H1 and P-H2 for the hinge, and P-D1 to P-D3 for the sampling depth. The population was amended once, from seven pairs to eleven, with the four added pairs chosen on availability and structure alone and before any certificate on them was computed; no criterion was altered or retired, and every one was re-adjudicated on the full corpus. The Theorem 1 that P-M1 names is the runner-up program of Prop. 2, which Theorem 1 extends; the stress run adjudicates P-M1 on both, and records where the runner-up program is silent. Two hold on the original seven and fail on the eleven. The registered criterion asked for the aggregate certificate to be strictly smaller than the per-task one on every board. On the mean-outcome scale the script compares them on, it is not: on two pairs, among them the one carrying the highest correlation in the corpus, the per-task decision is the better constrained of the two. On the per-task scale a third pair ties. The registered criterion asked for no board to reach a fresh read that selects within the tolerance with probability .99 at its own stated depth. One does: of the four registered grids, the fifty-trial one resolves its own decision, and the other three do not. Both declarations are written, word for word, by the scripts that adjudicate the criteria. Table 1 and App. A report the first finding and Sec. 4 the second. Nothing reported beyond what the criteria name was registered; among the rest are the permutation class of Prop. 11, behind Table 4 and the triangles of Fig. 1, and twelve further quantities named here. Three are the exact binomial box beside the registered plug-in one, three structured board laws beside the registered uniform one, and the pricing of every pair at the task-averaged summary's own exchange rate beside the pooled one (Prop. 10). A fourth is the shortlist decision of Sec. 2, which at $k = 1$ is the registered aggregate one, reported with the smallest shortlist each correlation certifies. A fifth is the aggregate bar of Cor. 3, which prices the registered aggregate decision the way Sec. 3 prices the per-task one. A sixth is the board-free population of App. C, one source carrying twelve distinct observations. Two more are the null of App. A, which all three summary comparisons are read against, and the count of pairs each summary resolves outright. Two more are the between-task share of Cor. 4 and the board-free form of a declared accuracy, both of App. A. The last two are what a correlation and a rank statistic certify together and what the recommended converters certify as a bundle. Every criterion is adjudicated on what its register fixed.

Checking the results. In the recorded environment `regenerate.py` requires the regenerated JSON artifacts to match exactly, and it compares every figure pixel by pixel at 200 dots per inch. `numeral_coverage.py` scans the numerals of two or more digits or with a decimal point, outside citations, cross-references, identifiers, dates and layout dimensions, and requires each to match a value a released script prints or an artifact holds, or a stated exemption. A match alone is weak evidence, since most such digit strings occur somewhere among the released values, so the central quantities are bound to the exact fields they report. The gate checks the rounding direction App. A states at every range endpoint and threshold figure it binds. A count spelled out as a word is invisible to that scan, and is checked only where a pin binds it to its field.

Cost. The whole pipeline runs in a little over an hour on one core, most of that in three Monte Carlo steps of `margin_audit.py`: the random-board null, the matched null and the scan of the

three structured laws. The exact programs behind Tables 1 and 2 take seconds, and the fast mode of `regenerate.py` skips only the stress run behind App. A's numerical verification. The registered depth criterion is adjudicated on a 100000-replication sampling simulation, whose depth search runs at ten times that replication count, because it compares a simulated probability with a threshold rather than reporting a value. The decision-safe probabilities Sec. 4 prints are not that simulation's estimates: they are the exact convolution of the per-task laws over the rationals, which is why a third decimal is printed, and the generator asserts on every run that the two agree. The depths Sec. 4 prints are the exact search's minima; the simulated search returns 73 on Google Robot where the exact one returns 72, and agrees on the other three boards. `code/requirements.txt` pins every third-party package the released code imports, though not the packages those depend on, to the version `code/data/build_env.json` records, and that record also names the Python release, the figure font and the TeX distribution the exact comparison assumes.

E Extended related work

Correlation as evidence for a decision. The question was first asked about personnel selection. Taylor & Russell (1939) tabulated the fraction of selected candidates who succeed against a test's validity coefficient, the base rate and the selection ratio. The model of Brogden, Cronbach and Gleser made the expected gain proportional to the coefficient (Cronbach & Gleser, 1965). Both are average-case under a bivariate normal model and one threshold decision. Tripuraneni et al. (2024) give the modern form for online experimentation, an exact orthant-probability map from proxy correlation to the probability of a wrong shipping decision. The surrogate-endpoint literature reached the same conclusion (Baker & Kramer, 2003; Fleming & DeMets, 1996), from Prentice (1989)'s criteria to the threshold effect of Burzykowski & Buyse (2006). The proxy-reward line is separate: Laidlaw et al. (2025) measure a proxy's correlation with the true reward under a reference policy's occupancy measure, and define hacking as a policy optimised for that proxy doing worse than the reference on the true reward. Liu et al. (2026) maximise the worst case over the rewards at that correlation with the proxy, solving the adversary in closed form: the same constraint, in population, over continuous policies. Xu et al. (2025) exhibit outcome correlation tending to one while the induced treatment rule inverts. Our contribution in this line is to replace a statement in population, under a model, by a maximum over a fixed printed table, which is attained and needs no model.

Decision-aware prediction. The predict-then-optimize framework (Elmachtoub & Grigas, 2022) studies exactly the composition we analyse: a prediction is made, a decision optimises against it, and the loss is that decision's regret. El Balghiti et al. (2023) introduce the distance from a prediction to the set where the induced decision is ambiguous, the multiclass margin on the simplex; our $d(a)$ is the squared distance to the complementary object, the cone producing a given decision. Hu et al. (2022) bound decision regret by a norm of the prediction error, with fast rates under a Tsybakov-style condition on the decision margin (Tsybakov, 2004). In that language $\varepsilon^\star(r)$ is the tight counterpart of their Lipschitz bound, with the budget read off a published correlation, and $\mu(\varepsilon)$ is a functional of their margin condition rather than an assumption about it. Minmax regret over an ellipsoidal uncertainty set is studied in combinatorial optimisation by Chassein & Goerigk (2017); (2) takes that worst case over a spherical set, for per-task selection. The closest two-sided statement is Skalse et al. (2024), whose standardised reward distance induces an upper and a lower bound on worst-case regret and is unique up to bilipschitz equivalence among distances that do. That is this paper's programme one level of abstraction up, over reward functions and policies rather than over a printed table and a taskwise argmax, and its tightness is tightness as a class: $\varepsilon^\star(r)$ is a single attained maximum, at the statistic benchmarks actually publish. Karwowski et al. (2024) supply the matching geometry, an angle between proxy and true reward, Goodharting growing rarer as it shrinks; (2) is that picture made quantitative on a fixed table, with the angle read off a published correlation. Maximum regret is a classical criterion for choice (Manski, 2004), and under interval data its worst scenario is classical too (Kouvelis & Yu, 1997). Aissi et al. (2009, Prop. 7) put the chosen elements at their worst bounds and the rest at their best, which (9) and Lem. 2 specialise to per-task selection. New here are the margin form (10), the attained largest radius, and the composition with per-cell binomial intervals.

Limits of proxy evaluation. Dorner et al. (2025) prove that a judge with agreement rate r can be trusted to rank two models only when their true scores differ by more than $2(1-r)$, and that the requirement does not soften as the evaluation set grows. Our Proposition 2 is the per-task-selection

analogue, with a constant and a rate rather than an asymptotic statement. Wen et al. (2025) show reward models of similar accuracy inducing very different policies, and Kocmi et al. (2021) reframe metric validation as decision accuracy rather than correlation. Landesberg (2026) proves that a judge's recovery rate is not identified by its correlation alone, the population statement whose fixed-table counterpart is Sec. 2. Lambert et al. (2020) named objective mismatch; Yu et al. (2026) make the same diagnosis for world-model evaluation in prose, and Zhang et al. (2026a) score world models by closed-loop task success instead of visual quality. Badithela et al. (2026) answer the question this paper leaves aside: from a few paired real and simulated evaluations, treated as a prediction-powered inference problem, they give confidence intervals on real policy performance, correcting the simulator's bias rather than pricing its summary. That needs paired data and a sampling model; every certificate here needs only published numbers, and bounds what the summary licenses rather than what the policy does. A benchmark with paired data should prefer theirs. The same correction exists for language-model judges, for scores (Boyeau et al., 2025) and for rankings (Chatzi et al., 2024), and Kocmi et al. (2024) map a metric's score difference to the accuracy of the pairwise decision it implies. Two sources in our corpus measure an automatic evaluator against a human reference: Lù et al. (2025) report judge and checker precision and recall against expert annotation, and Zhou et al. (2025) the agreement between an autonomous real-robot protocol and a human-supervised one. Both measure the evaluator's agreement with its reference on its own scale, and both stop where this paper starts: at what such a summary licenses about the choice, on the scale of the outcome.

Rearranging an observed matrix. Proposition 11 permutes within the margins of an observed matrix to extremise a functional of it, which is the operation the Fréchet-class literature calls the rearrangement algorithm, on the line from Hardy, Littlewood and Pólya (Puccetti & Rüschendorf, 2012; Embrechts et al., 2013). There it is a heuristic, because the row sums couple every column and the iteration can stall at a local arrangement. Here the coupling collapses: the correlation depends on the permutations only through the single additive scalar C, so the maximisation separates over tasks and is solved exactly rather than approached.

Benchmark size and reliability. A reader may expect size to help, on the strength of the prophecy formula of Spearman and Brown (Spearman, 1910; Brown, 1910). That formula governs *reliability*, the attenuation of a correlation by measurement noise, which does rise with test length; the bar is a *required validity* threshold on two fixed score vectors, with no sampling model, and Prop. 8 shows it converges rather than falls. The two act in opposite directions on the practitioner's decision, which is why App. B states the attenuation caveat. Maia Polo et al. (2024) find a small fraction of items reproduces most of a ranking, the design-effect literature records the same diminishing return (Kish, 1965), and Vincent et al. (2024) ask how many rollouts a trustworthy evaluation needs.

Reporting practice. Miller (2024) sets out the current standard (question-level statistics, clustered standard errors, power analysis), and Card et al. (2020) document how underpowered typical comparisons are. Kotawala (2026) reports a per-pair resolution ratio, the sample size a leaderboard pair realised over the size it would have needed, the diagnostic closest in purpose to the margin profile we recommend. Gordienko et al. (2026) formalise the cheapest ranking-changing perturbation as a social-choice problem, an all-or-nothing cost model where ours is quadratic (Xia, 2012); the cheapest perturbation that changes a decision is the same question a classifier's minimal adversarial distance asks (Moosavi-Dezfooli et al., 2016). Fu et al. (2021) print regret@k beside rank correlation, realised rather than worst case, and Wang et al. (2026) report MMRV beside Pearson without converting it; Prop. 3 supplies the factor. Publishing the margin because it sets the verification effort is established in risk-limiting audits (Lindeman & Stark, 2012), and Cor. 1 and Prop. 4 say why those numbers are the ones worth printing.