COMPARING MODELS FROM INCOMPLETE EVALUATION EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Leaderboard tables assert an ordering over all models, but the evidence is incomplete: models are evaluated on different benchmark subsets. We ask which of those comparisons the observed evidence establishes. Taking the platform's declared model universe, benchmark universe and aggregation operator as given, we add one counterfactual estimand: that operator on the matrix *completed* over that universe, unobserved cells held to admissible ranges. Auditing the reported aggregate would be tautological — it is that operator's own output. Each comparison is labelled established, possible, refuted or undetermined. For HELM's mean win rate, pairwise claims admit exact closed-form bounds under the pinned rule reconstruction, to which established leaderhood reduces; possible leaderhood is genuinely joint. On HELM's `classic: core_scenarios` Accuracy universe (81.7% observed), 4 of the 45 reported top-ten orderings are established and no model leads in every completion; two further declared universes establish 0 and 3. Complete cases decide all 45 because a measured block leaves nothing to decide, while the two deletion remedies cost the same evidence but name different leaders. Masking a real block, the pattern of missingness, not its amount, changes what is established. The audit is also a decision problem, and one step of it is solved exactly: the measurement that settles the most reported claims *whatever* it returns is computable from the published matrix alone, is provably blind to the value it buys, and certifies in advance how many claims a budget will settle.

1 INTRODUCTION

Leaderboard summaries are how the field judges progress and chooses baselines, and each is read as a set of definite comparisons: this model is ahead of that one, these are the ten best. Those comparisons aggregate measured scores, and the measurements behind them are usually incomplete.

The leaderboard we study presents a total order over 83 models, computed from an 83×207 model by benchmark matrix that was never filled in: models were submitted to different benchmark subsets by different groups. In the published HELM matrix (Liang et al., 2022), 56.9% of the cells are observed, and not at random — they occupy only 62 distinct per-model coverage sets, with per-model coverage from 8.7% to 89.4% — so a reader cannot tell which reported comparisons rest on shared evidence and which on two disjoint portions of it.

Two standard remedies answer a different question. Imputation and matrix completion replace an unknown score with a presumed value (Zeng & Papailiopoulos, 2026; Himmi et al., 2024); deletion restricts to a completely measured submatrix, dropping either the models that missed a task (Wang, 2026) or the tasks not every model ran. Both work, and neither answers the leaderboard's question: on the declared universe of §4.1 each leaves a completely measured block on which all 45 reported top-ten orderings are decided, at almost the same evidence cost (53.5% and 53.0% of the observed cells) and with *different* leaders, neither of them necessarily the one the leaderboard crowns. Certainty about a small question, bought by discarding half the evidence, does not transfer back (§4.1, Appendix E).

This paper takes the alternative route: it keeps both the evidence and the rule fixed and asks which reported comparisons are actually *established* by the evaluations observed. The rule is the platform's own, used as published; what we supply is the *estimand* — the platform's rule read on a *completed* declared universe rather than on the partial evidence it holds. That reading is a counterfactual, stated

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.

as one: it asks whether a reported comparison is robust to filling in the evaluations the platform's own declaration already covers. Auditing the reported aggregate directly would be circular: the published ordering *is* that aggregate's output on the observed entries. Each unobserved cell is constrained only by an admissible range, and a claim is *established* when it holds in every completion consistent with those ranges, *possible* when it holds in at least one, and *refuted* when it holds in none. On HELM's declared `classic: core_scenarios` Accuracy universe, 4 of the 45 orderings inside the reported top ten are established and *none* is refuted: the evidence is silent about most of what the leaderboard asserts rather than contradicting it. Restricting to the scenarios every declared model ran reports all 45 as established, having discarded 9 of the 16.

The 83×207 matrix is the motivation, not the audited object: auditing all 207 columns would need a declared rule for each, and the release's own declared range contradicts the metric for the bits-per-byte ones (Appendix F). The audit therefore uses the platform's declared `classic: core_scenarios` Accuracy universe — 67 models, 16 scenarios, 81.7% observed — all three platform-supplied; the completed-universe estimand is ours, and the only thing here we chose. §4.2 and §4.3 re-use a second, completely measured HELM slice as a controlled testbed, and every empirical claim is scoped to these two slices and to the mean-win-rate rule rather than to leaderboards generally (§4.5).

Contributions.

1. **Claim-level identification for incomplete evaluation evidence.** A partially observed matrix under a platform's *declared* rule, each unobserved cell held only to an admissible range, with the assumption level an explicit axis reported at every level rather than fixed by us (§3.1–§3.2).
2. **Exact computation for a rule platforms actually use.** For HELM's mean win rate the pairwise identified set is a sum of per-column extremes, in closed form and with no optimiser; established *leaderhood* reduces exactly to those tests (Proposition 1), while *possible* leaderhood is genuinely joint — per-rival feasibility does not imply a common completion (Proposition 2) — and is decided by a MILP whose witnesses we certify by re-scoring under the declared rule (§3.3).
3. **What incomplete evidence supports in practice.** An audit of a real platform's claims under its own rule, run on three of that platform's declared universes (§4.1, Appendix D); the common-support contrast that prices the deletion remedy; a masking study in which the *pattern* of missingness, not only its rate, is the experimental variable (§4.2); and a joint-versus-pairwise stress test on real measurements that finds strict gaps (§4.3).
4. **A measurement rule for the claims the evidence leaves open.** The identification analysis makes the follow-up decidable, and this one is deployable: the cell that settles the most reported claims *whatever* its measurement returns is exactly computable from the published matrix, the mask, the declared rule and the admissible ranges; it never reads the value it is about to buy, proved mechanically rather than asserted (§J); and it carries a lower bound, known before anything is bought, on how many claims a budget will settle (Proposition 7). The bound is unconditional on the declared metric domain, and *per segment* on a range derived from the observed data, where §J measures how often that condition is falsified rather than assuming it holds. On the block where the truth is known the rule settles more than any heuristic the earlier version of this study compared — an advantage §J separates into the part the bound provides and the part it does not (§4.4).

2 Related Work

Partial identification. That data may identify a *set* rather than a point, reported as a function of what one is willing to assume, is classical in econometrics (Manski, 1990; 2003; Tamer, 2010); our hierarchy is that program applied to a new object. Theirs must do inference because the object is a population parameter estimated from a sample, whereas ours are sharp by construction, about *this* matrix and about *which model leads* rather than a scalar — where the quantifier order of §3.3 matters.

Necessary and possible winners. Computational social choice has studied what partial preference data determines since Konczak & Lang (2005), with complexity results for top-k winners under scoring rules (Imber & Kimelfeld, 2020b;a; Xia & Conitzer, 2011) and for partial tournaments (Aziz et al., 2015); the established/possible dichotomy and the factorization behind Proposition 1 are theirs. The data object differs — a missing *rank* or match outcome there, a missing *benchmark score* here — so the unknown lies in an interval, the constraint is a box, and the aggregate is the declared rule composed with per-column order statistics, redone per rule because each column's comparison set is the models evaluated in it.

Elicitation and active evaluation. Which comparison to buy next is the query-selection problem of preference elicitation, where the aim is to call the winner with as few queries as possible (Pini et al., 2007; Lu & Boutilier, 2020; Shah & Yu, 2026), and the same move appears in evaluation, where the next system pair (Mohankumar & Khapra, 2022), the next benchmark item (Li et al., 2025) or the decision to stop (Arviv et al., 2026) is chosen rather than fixed. Three things differ here. Their unknown is a preference relation or an unobserved item outcome that the query reveals; ours is a score constrained to a declared interval, so an acquisition buys a *narrowing of an identified set* rather than the observation of a preference. Their objective is a decision — the winner, the ranking, the stopping point; ours is the count of *already published* claims the evidence comes to settle, a set of predicates on the completion, scored in the worst case over it. And because the aggregate is a column-wise mean that worst case is a finite list, so the score is exact and in closed form (§4.4) — which the general elicitation problem does not offer.

Ranking uncertainty and estimand-aware comparison. A distinct line quantifies *sampling* uncertainty — rank intervals and confidence sets for ranks (Neuhof & Benjamini, 2026b;a; Mogstad et al., 2024) — on a complete matrix, shrinking with more test items; ours requires the missing *evaluations*. Assembling a *ranking* conclusion from marginal pairwise comparisons is unsound unless the comparisons are simultaneous; our gap is that failure transposed to unobserved cells, where it is a provable, one-directional wrong label. Closer to us, Huang (2026) shows that a leaderboard comparison's meaning depends on the estimand, and proposes a common-support check within a benchmark; we extend the object from within-benchmark estimand choices, where the data are complete, to across-benchmark coverage.

Imputation, deletion, and coverage. Matrix completion predicts missing scores (Zeng & Papailiopoulos, 2026), and Himmi et al. (2024) repair a leaderboard by imputing a partial ranking and aggregating with Borda count; both convert an unknown into a presumed value, whereas we keep the set of completions the evidence admits. Wang (2026) give a stereological theory of intrinsic ambiguity that persists with infinite data per benchmark, and entry-specific bounds argue a smaller submatrix can be preferable for *estimation* (Xi et al., 2024); what a complete-case restriction costs on a real platform is in Appendix K.

3 Problem Formulation and Method

3.1 Problem Setup

What the platform supplies, and the estimand we add. A platform supplies three things: a declared model universe M, a declared benchmark universe B, and an aggregation rule ρ mapping a complete score matrix to one aggregate per model. We take all three from the platform. What we add is the *estimand*. The platform applies its rule to the entries it happens to hold; we apply that same rule to the matrix *completed over the full declared universe* B, and label claims about the result. That extension is ours and counterfactual: it asks what the platform's own published rule would report if the declared universe were filled in. The leaderboard asserts $A > B$ from the entries it has; we ask whether that assertion survives completion of the universe the leaderboard itself declares. A claim that fails is one the platform's declaration does not support — sharper than saying the platform got it wrong.

Why the reported aggregate is not the thing to audit. Write ρ_{rep} for the rule as the platform actually applies it, on the observed entries $\Omega \subseteq M \times B$, each model averaged over *its own* present columns. Auditing ρ_{rep} is tautological: the published order *is* its output on Ω, so there is nothing

to establish or refute, only to restate. The identification question lives entirely in the step from ρ_{rep} to ρ, and ρ is defined on the complete matrix, so that step is a choice about the unobserved cells — where the assumption hierarchy of §3.2 enters. The labels answer whether a comparison the platform reports as definite stays definite on that basis.

The rule, pinned to source. HELM's leaderboard aggregate is the *mean win rate*, and we use the upstream implementation rather than a paraphrase: in `stanford-crfm/helm`, `summarize.py`'s `compute_aggregate_row_win_rates` computes, for each column b with at least two present values,

$$v_2 w_b(v) = \#\{i : S[i,b] < v\} + \tfrac{1}{2}\big(\#\{i : S[i,b] = v\} - 1\big), \qquad \mathrm{wr}[i,b] = \frac{v_2 w_b(S[i,b])}{n_b - 1}, \quad (1)$$

with $n_b \le |M|$ the number of models present in column b *as reported* — on the complete universe of §3.3 every column has $n_b = |M|$, which is what makes the denominator below a constant — and then $\rho(S)[i] = \mathrm{mean}_{b\,:\,S[i,b]\ \text{present}}\mathrm{wr}[i,b]$. A model's aggregate is thus a mean over *its own* present columns against *that column's* present models — ρ and ρ_{rep} at once, differing only in the matrix each is evaluated on.

The object of every result below is the pinned reconstruction of HELM's declared rule *as a component of an estimand of our construction*: exactness claims below are conditional on it. The operator reproduces HELM's published aggregate to a maximum residual of 0.019 win-rate units (Appendix C), which also bounds what a discrepancy of that size can do to the labels: exact under a pinned reconstruction is not an exact audit of the published column, and where the two could diverge we report the bound rather than assume agreement.

Admissible ranges. For each $b \in B$ we fix a stated range $A_b \subseteq \mathbb{R}$, the set of values an unobserved score for benchmark b could take, and the feasible set is $F = \{\, S' \in \mathbb{R}^{|M|\times|B|} : S'[i,b] = S[i,b]$ for $(i,b) \in \Omega$, and $S'[i,b] \in A_b$ otherwise $\}$, a product over columns, $F = \prod_{b\in B} F_b$ — the structural fact §3.3 exploits. A_b states what is *not ruled out*, never a belief about what the value is.

Claim labels. A comparison claim is a predicate c on complete matrices — for example "$\rho(S')[i] \ge \rho(S')[j]$" or "$i \in \arg\max \rho(S')$". We label it *established* if $\forall S' \in F,\ c(S')$, *refuted* if $\forall S' \in F,\ \neg c(S')$, and *possible* if $\exists S' \in F,\ c(S')$; a claim that is neither we call *undetermined*. Refuted is the dual of established under negation, so we track established and possible and derive the rest; Figure 7 (Appendix L) draws the whole construction in one panel.

3.2 Assumptions on Missing Evaluations

Assumption hierarchy. The choice of A_b is the only place a substantive assumption enters, so we make it an explicit axis and report every result at every level: A0 takes the benchmark's declared metric range ($[0, 100]$ for a percent metric), A1 the observed span $[\min_b, \max_b]$ of column b, and A2 that span cut to the half around its median, $A2_b$ of equation 12. A0 $\supset$ A1 $\supset$ A2 orders the feasible sets, which makes the axis a soundness/completeness trade rather than a free tightening: a weaker assumption identifies fewer established claims but admits more completions a reader would reject, and a stronger one can *exclude the truth*, because the extreme value that would settle a narrow claim may be exactly the unobserved cell.

A0 is the weakest of the three — the widest feasible set, and the only level not built from the observed data — but weakest is not none: it *asserts* that the release's declared domain for a metric is the set that metric can take, so it applies only where the platform declares a range that describes the metric. Three of §4.2's columns are bits-per-byte metrics whose recorded percent range does not describe them, so that study is reported at A1 (Appendix F); §J adjudicates the declaration column by column. A1 is the level a practitioner is most likely to reach for, and its convenience hides a defect: built from column b's *observed* values, it excludes a hidden score outside the observed support, making A1 unsound in principle rather than conservative. §4.2 measures how often that happens.

3.3 PAIRWISE AND JOINT CLAIMS

Pairwise claims decompose exactly. The target universe is complete by construction, so every model's aggregate is a mean over the same $|B|$ columns and $\rho(S')[i]-\rho(S')[j] = \frac{1}{|B|}\sum_b\big(\mathrm{wr}[i,b]-\mathrm{wr}[j,b]\big)$. Each summand depends on column b alone, and F is a product over columns, so the supremum over F separates:

$$\sup_{S'\in F}\big(\rho(S')[i]-\rho(S')[j]\big) = \frac{1}{|B|}\sum_{b\in B}\sup_{S'[:,b]\in F_b}\big(\mathrm{wr}[i,b]-\mathrm{wr}[j,b]\big), \tag{2}$$

and likewise for the infimum. Equation equation 2 reduces a $|M|\times|B|$-dimensional optimisation to $|B|$ independent one-column problems, each in closed form: pinning the pair's own values at an extreme end of A_b is optimal, since the difference is monotone in each, and the remaining free cells are then set independently from a finite candidate set. Writing v_i, v_j for those pinned values and $[\ell,u]=A_b$, one further free cell contributes a piecewise-constant function with breakpoints exactly at v_i and v_j, so its extreme over $[\ell,u]$ is attained on $\{\ell,u,\frac{\ell+u}{2}\}\cup(\{v_i,v_j,\frac{v_i+v_j}{2}\}\cap[\ell,u])\cup\{\frac{v+w}{2}:v\in\{v_i,v_j\},\ w\in\{\ell,u\}\}$. The between-breakpoint representatives are what make this sound: strictly between v_i and v_j the contribution is constant at its maximum $+1$, which a breakpoint-only search misses — at $v_i=0.7$, $v_j=0.3$, $[\ell,u]=[0,1]$ it returns $\frac{1}{2}$ against a true maximum of 1. Appendix B checks the implemented set against exact region enumeration on 4.6×10^5 adversarial instances, including $v_i=v_j$, with no mismatch.

Leader claims do not. Deciding whether a model can lead in *some* completion differs from deciding it against each rival separately, and the two are not interchangeable:

Proposition 1 (Established leaderhood reduces exactly)**.** *i is the leader in every completion of F if and only if, for every rival j, $\inf_{S'\in F}\big(\rho(S')[i]-\rho(S')[j]\big)\geq 0$.*

Proposition 2 (Possible leaderhood does not reduce)**.** *The statement $\exists S'\in F\,\forall j:\rho(S')[i]\geq\rho(S')[j]$ is strictly stronger than the per-rival relaxation $\forall j\,\exists S'\in F:\rho(S')[i]\geq\rho(S')[j]$.*

The relaxation is necessary but not sufficient: the completions that defeat rival j and those that defeat rival k need not coincide, and that is what fixes the polarity of every error the screen can make.

Proposition 3 (The relaxation screens one-sided)**.** *Let $\mathrm{marg}(i)=\min_j\sup_{S'\in F}\big(\rho(S')[i]-\rho(S')[j]\big)$. If $\mathrm{marg}(i)<0$ then no completion of F makes i the leader. Consequently a row with $\mathrm{marg}(i)\geq 0$ can fail to be joint-possible — and when it does, the relaxation is wrong by asserting a leader — but a row the screen excludes is never a false negative, by construction rather than by measurement.*

Proofs are in Appendix A; the asymmetry is what §4.3 measures, and the screen errs only by admitting too much.

Certifying the joint statement. Deciding it requires an optimiser: encoding each column's ranking with integer variables makes the win rates linear in them, giving a MILP over F projected onto the aggregate. *The solver proposes, the declared rule disposes*: we rebuild any proposed completion from the solver's own *integral* relation, re-score it with the declared rule, and accept it as a witness only if it reproduces the claim. The encoding, its tie treatment and tolerances, and the status criterion under which a negative answer certifies that no completion exists are in Appendix H; it is validated against exhaustive enumeration of F (`tests/test_joint.py`, Appendix B).

3.4 EXPERIMENTAL SETUP

(a) A naturally incomplete platform matrix. HELM's declared `classic: core_scenarios` Accuracy universe — 67 models, 16 scenarios, and its own aggregate column `Overall (classic: core_scenarios)/Accuracy/Mean win rate` as the rule, none of the three chosen by us — leaves $67\times16=1072$ cells, 81.7% measured. Score recovery and the direction convention are recorded with the audit (Appendices C and F).

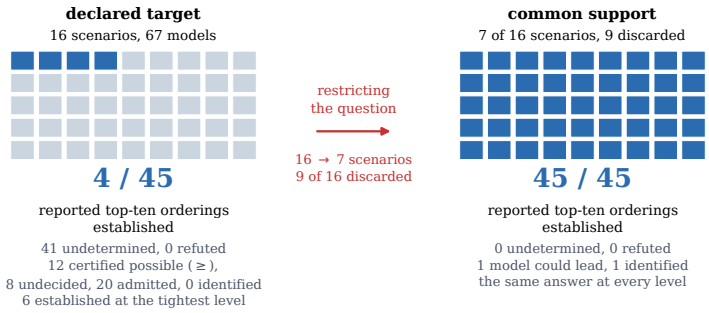

Figure 1: The paper's headline result and the remedy it prices. Each square is one of the 45 pairwise orderings inside HELM's reported top ten, at the weakest assumption level (A0). Left: on the declared 16-scenario universe the observed evaluations establish 4 of them and refute none. Right: restricting to the 7 scenarios every declared model ran — discarding 9 of the 16 — establishes all 45. The complete-case answer is an answer to a smaller question.

(b) A completely measured real block, masked. Ground truth needs a complete matrix. The masked base is HELM's `classic: targeted_evaluations` Accuracy universe restricted to the 32 models covering all 22 declared scenarios: all 704 cells are real published measurements, so masking hides measurements and never invents them. Four hiding patterns — **mcar**, **budget**, **block**, **column** — calibrated from the published matrix's own coverage statistics (Appendix G), give 280 masked instances (4 patterns $\times$ 7 rates $\times$ 10 seeds), each analysed at A1 over all 45 pairwise orderings inside the *true* top ten.

(c) A joint-versus-pairwise stress test. The same block is masked into sub-universes under the block pattern at five missingness rates: declared sizes $|M| \in \{5, 6, 8, 10\}$ at A1 and $\{5, 6, 8\}$ at A2, 350 configurations and 2,400 screened leader queries in all. Every model goes to the cheap screen, and every model it does not exclude to the exact solve. No $n = 12$ cell was run, so every incidence in §4.3 is conditional on the sizes actually sampled (Appendix I).

(d) Acquisition. The same block supports a third experiment, and unlike the first two it is a decision problem: an episode in which each rule may buy 20 of the 704 cells, seeing the observed matrix, the mask, the declared rule and the ranges, and learning a cell's value only by buying it. Ten rules are compared — six computable from the published matrix alone, three heuristics kept from the earlier version of this study, and an outcome-aware one-step greedy comparator that reads the value it is about to acquire and is therefore not deployable (Appendix J).

4 Experiments and Results

4.1 What Does the Real Leaderboard Evidence Establish?

HELM's reported order interleaves models scored on different benchmark sets: its 67 models fall into seven coverage groups (sizes $29, 25, 9, 1, 1, 1, 1$), and the reported leader `Llama 2 (70B)` is averaged over *10* of the 16 scenarios while three of the five models immediately below it are averaged over all 16, each normalised against a different population. Figure 1 draws the headline result beside the remedy it prices, and Table 1 (Appendix E) records every label. Three things follow.

The evidence is silent, not misleading. 41 of the 45 orderings are undetermined at A0 and A1 and 39 at A2, while *zero* are refuted at any level: the observed evaluations never establish the opposite of a reported ordering. Refutation is not excluded by construction — it needs $\inf_{S' \in F}(\rho(S')[i] - \rho(S')[j]) > 0$, a live possibility at 81.7% coverage — so that it never fires is a finding rather than a definitional artefact.

No model leads in every completion. The identified-leader count is zero at all three levels, and the reported leader's worst-case margin is negative at each (-0.354 at A0, -0.348 at A1, -0.308 at A2). The exact joint solve certifies 12 models able to lead at A0, 14 at A1 and 8 at A2, where the cheap per-rival relaxation would admit 20, 16 and 9; the certified counts are *lower* bounds and the

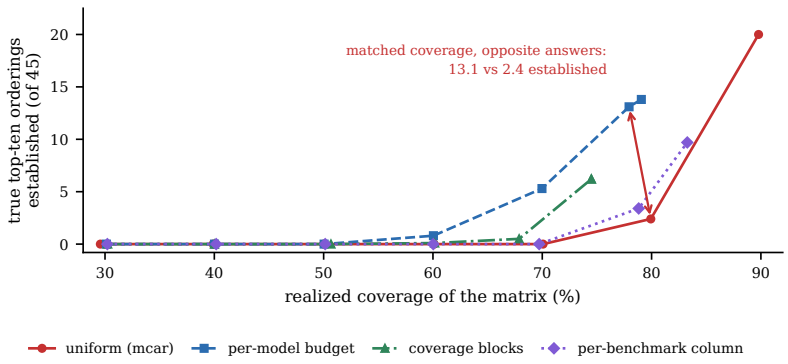

Figure 2: Identification against the amount of evidence, under four hiding patterns, on a completely measured real block (means over 10 seeds, at A1). Observed evidence is held *comparable* rather than equal — the calibration meets its coverage tolerance in 19 of 28 settings (Appendix G), so the comparison is not at identical coverage. Two points *less* complete evidence establishes five times as many true top-ten orderings. How much is missing is not a sufficient statistic for what the evidence settles.

excess *upper* bounds, because the solver left 8, 2 and 1 rows undecided (Table 1), so at A0 the true count lies in 12–20. Exactness is claimed only where it was achieved: the established labels and the zero identified leaders rest on closed-form bounds, exact at every level.

Complete cases answer a smaller question. Restricting to common support — the 7 scenarios *every* declared model ran — discards 9 of the 16 declared scenarios and 407 of 876 observed cells, leaving a completely measured 67×7 block. There F is a singleton, so every ordering is decided by direct computation: "all 45 established" is an identity rather than a finding, the uncertainty *was* the missing cells. The contrast buys no *power*; it changes the *question*, from which comparisons the declared evidence settles to which hold on a smaller evaluation. What is genuinely reportable is the part we did not assume: the reported leader `Llama 2 (70B)` survives the restriction, by $+0.0152$ over its runner-up.

The other deletion remedy behaves the same way, and the two are not interchangeable: dropping models rather than scenarios — the restriction Wang (2026) use — leaves an equally complete 29×16 block, all 45 established, at the same evidence cost (53.0% against 53.5%), but crowns a *different* leader, `text-davinci-002`. Two remedies of equal cost, disagreeing about the answer: that is why this paper fixes the evidence and the rule rather than adopting either (Appendix E).

4.2 Does Missingness Structure Matter Beyond Its Amount?

A naturally incomplete matrix has no ground truth: an ordering can be labelled undetermined but not checked against what complete evidence would have said, and the masked block supplies that check.

The pattern changes the answer at nearly matched coverage. At 79.9% realized coverage uniform independent deletion leaves 2.4 of the 45 true top-ten orderings established, while at 77.9% — two points *less* complete — the budget pattern leaves 13.1 (Figure 2), a factor of five and a half; the block pattern repeats it (Table 4). The mechanism we offer as an *explanation*, not a demonstrated general result, is concentration: independent deletion scatters missing cells across every row and column, whereas structured missingness concentrates the gaps into large fully observed submatrices, so pairs with enough shared scenarios are decided and the rest by nothing.

The labels are conservative, and silence is their cost. Across all 280 analyses zero true top-ten orderings are *reversed*, and the true leader is a possible leader in every one: the analysis never excludes the model we know to be correct. That zero is a measurement rather than a guarantee: the true completion lies inside A1's feasible set in 1 of the 280, and in 277 of the other 279 the observed span excludes a hidden value in a true-top-ten row. The identified-leader count is zero everywhere, including at 89.8% coverage where 20.0 of the 45 orderings are established — the leader claim requires *all* $|M| - 1$ comparisons at once, and any missingness reopens at least one.

In this experiment the platform's own answer is wrong about a third of the time: the model the native rule reports as leader is the true leader in 188 of the 280 masked instances, while our labels are never reversed. The count is pooled with no interval — the instances share a base block, a mask family and a seed set — and does not transfer to §4.1, which has no ground truth.

4.3 When Does Pairwise Feasibility Fail for Leader Claims?

Proposition 2 says the cheap per-rival test is not sufficient for possible leadership, not by how much. Raising a model's value in a free cell weakly increases its own win rate and weakly decreases every rival's, so one row's free cells are a *monotone instrument*: a gap requires free cells in *rival* rows. We screen at $k = 1$ because the relaxation $\min_j \sup_F$ does not depend on k, while the exact top-k feasible set grows with it: absence of a gap at $k = 1$ rules out gaps at every k (§I gives the nesting). A0 was not run in the grid, so every count here is an A1 or A2 count. Of the $2{,}400$ leader queries the screen excludes 497 as provably impossible, the exact solve decides $1{,}811$ of the remaining $1{,}903$, and 92 are time-limited with no verdict (Table 5, Appendix I).

Of the $1{,}798$ decided rows that admit a certified witness, the cheap test and the joint solve agree; the thirteen that disagree are the point. All run in one direction — the cheap test records a leader no completion of the evidence can realise — which is Proposition 3 rather than a second finding: manufactured certainty, not doubt, the same failure mode the common-support restriction exhibits. Their slack runs from 0.0077 to 0.070 win-rate units: short of a few percentage points of aggregate win rate, a model can be certified a leader it cannot be. The count is a *lower* bound — a time-limited row could add to it and cannot remove from it.

Six of the thirteen sit on the relaxation's boundary: they have $\min_j \sup_F(\rho(r) - \rho(j)) = +0.0000$ exactly, the relaxation's best completion leaving r *tied* with a rival. Under the permissive reading the screen uses — admit a row when the cheap bound is ≥ 0 — all thirteen are gaps; under a strict reading, which counts only a *strictly* positive cheap bound, seven are, the other six being ties the relaxation cannot separate from non-gaps. We report **thirteen under the permissive rule and seven under the strict one**, and the rate $13/1{,}811 = 0.72\%$ of decided admitted rows as descriptive only, those rows being no random sample (92 of $1{,}903$ undecided). What the experiment establishes is *inexactness*, not prevalence (Table 7, Appendix I).

4.4 Which Measurement to Buy Next

The audit says which published comparisons the evidence leaves open, and a platform can close one by running the evaluation it never ran; which to buy is decidable from the evidence the platform already publishes. With $A(F)$ the number of target claims — here the 45 orderings inside the reported top ten — that F still leaves undetermined, the cell to buy is the one maximising the worst case $G(q) = A(F) - \max_{y \in A_q} A(F(q{=}y))$: the claims that settle *whatever* the measurement returns. The score is exact, in closed form, and reads only the published matrix, the mask, the declared rule and the ranges — never the value it is about to buy, established mechanically by a checker; ties among the maximisers go to the guaranteed contraction, so the proposed rule is the lexicographic (G, C) of §J and uses no distribution over outcomes.

Because the ranges are frozen for the episode the count cannot rise within a segment, so $G(q) \geq 0$ and the gains accumulate into a certificate held *before* anything is bought (Proposition 7). On the declared metric domain no measurement can fall outside the range, so the inequality there is unconditional — but that declaration is a property of the metric, and this block does not supply it for every column: three are bits-per-byte metrics, non-negative by definition with no maximum the release declares, so A0 is claimed on the 19 columns the release bounds and those three are outside the A0 universe (§J). On those 19, **zero** of the acquired outcomes left the declared domain. At A1, a statistic of the observed cells that a reveal can falsify, the inequality is stated *per segment*, restarts at each falsification and carries the falsifications as an explicit correction term; §J reports the counts rather than assuming there are none. Either way the guarantee is small — $\sum_t G(q_t) = 42$ against a realised drop of 123 — so §J separates what it contributes from what the tie-break does.

On the fully observed block, where every withheld value can be revealed and every claim checked against the truth, the rule establishes 19.8 of the target orderings in 20 acquisitions — 2.8% of the matrix — against 17.8 for the widest-pair heuristic of the earlier version of this study, 15.5 for the

column-completing one and 13.0 for random (Table 11). The two keys are *not* equally responsible: §J puts $+1.83$ of the $+6.83$ over random on the certified primary key and the rest on the tie-break. Ranking by the *expected* number of claims settled under a declared prior instead reaches 21.6, and the outcome-aware one-step greedy comparator 21.8; the comparator is optimal only at the step it stands on, and the expected-gain rule beats it in 6 of the 12 runs, which no episode bound would allow. Neither objective dominates; §4.5 says what that does and does not support.

4.5 Implications, Sensitivity, and Scope

What we claim, and what we do not. Not that leaderboards are wrong: the audit found *zero* refuted orderings. We claim that the evidence is silent about most of what a leaderboard asserts, that this silence is characterisable exactly and cheaply under a stated assumption, and that restricting to complete cases converts silence into confidence about a smaller question.

Acquisition is the diagnostic read forward. Because a platform could settle an undetermined claim by running the missing measurement, the audit's output is a decision problem as well as a list of open claims, and §4.4 solves one step of it exactly. That study supports no claim about the ordering of rules — nine designed instances, objectives trading mean against worst case, and an outcome-aware arm that is a one-step greedy comparator, not an episode bound — nor does it support attributing the rule's advantage to its guarantee, which §J shows is a certification layer while most of the realised gain comes from the tie-break. The rule itself is a certified one-step policy, unconditional on a declared metric domain and valid per segment on a range derived from the observed data (§J, §J); it is not an evaluated planner, and no identification result depends on it.

Limitations. Our structural results hold for HELM's mean win rate, a per-column rank statistic; the decomposition applies to any rule that is a column-wise mean of a function of that column alone, but not to rules with genuine cross-column coupling, such as Elo-style updates. That generality is *proved*, not demonstrated; the empirical scope is narrower — one platform, one operator, one historical release. Within it the one ingredient the data could vary, the declared benchmark universe, replicates the headline across three of HELM's own declared universes (Appendix D); no second platform or operator is present, so the mathematics carries the cross-platform claim and the demonstration only this one. The admissible ranges are boxes and deliberately do not encode the correlation of a model's scores across benchmarks, where imputation could enter legitimately as a labelled sensitivity assumption rather than a fact. The model universe is a snapshot: what should persist is the pattern of structured incompleteness, not the matrix.

Practical implication. A platform already computes its aggregate, and adding our labels costs only the closed-form pairwise bounds: report, alongside the ranking, how many adjacencies the assumption level establishes and which models could lead; §4.4 supplies the purchase order for what remains.

5 Conclusion

We have treated a partially observed model $\times$ benchmark matrix as the object and asked what the evidence establishes about the comparisons a leaderboard reports, studying the platform's own declared universes, operator and rule read on the matrix completed over the declared universe — the only ingredient we chose, and a counterfactual one: whether a reported $A > B$ survives completing the evaluation the platform declared and did not run.

For that estimand the answer is computable: pairwise claims in closed form, leaders by a certified MILP. On HELM's declared `classic: core_scenarios` Accuracy universe the evidence establishes 4 of the 45 orderings within the reported top ten and no model leads at any level; restricting to common support — the 7 scenarios every declared model ran — establishes all 45, discarding 9 of the 16. Two further HELM-declared universes, Robustness and Fairness, give 0 and 3 of 45, with no leader there either (Appendix D).

Incomplete evaluation evidence does not falsify a ranking; it limits which comparison claims the observed data establish. A measurement, not an argument, is what moves a claim across that gap.

AI USE STATEMENT

Generative AI tools were used throughout this work, and not only for assistance with presentation. They contributed to the research process itself: to developing the problem framing and the counterfactual estimand of §3.1, to the formal reasoning behind the propositions of §3.3 and the closed form equation 2, to the design of the experiments of §4.2–§4.3, to writing and debugging the experiment code and the test suites, to orchestrating the runs, to checking analyses, to preparing the figures, and to drafting and revising the manuscript, including revisions made in response to review. Where a sentence below says that a claim was checked, the check was executed with AI assistance as well; what makes it a check is that it is mechanical and reproducible from the repository, not who performed it.

Two things generative AI did not do. It did not produce experimental data: every score in this paper is a published measurement, masking hides real values and never invents one, and no synthetic or model-generated result appears anywhere. And it was not the sole authority for any claim that stands in the paper. Every number is emitted by a script in the project repository and is traceable to a stored experiment output; the closed-form bounds are checked against independent brute-force enumeration of the feasible set, and the MILP against exhaustive enumeration, in the test suites described in Appendix B; the citation record, the provenance of every dataset, and the derivation of each proposition were re-checked against primary sources; and the authors reviewed the work and take responsibility for its final content, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every number in this paper is produced by a script in the project repository, and the research ledger accompanying that repository records, for each dataset and each previous work cited, the provenance by which it was obtained. The score matrices are public downloads whose archive digests are recorded; masking hides real measurements and never invents one, so the ground truth used in §4.2 is a subset of published data. The declared aggregation rule is reimplemented from the platform's source and its agreement with the platform's own published aggregate is reported (§4.1) rather than assumed. Randomised procedures use recorded integer seeds; no result depends on dictionary or set iteration order.

ETHICS STATEMENT

This work analyses published benchmark results and does not involve human subjects, released model weights, or new model evaluations. Its subject matter is the interpretation of publicly reported comparisons between models, and its intended effect is to make the evidential basis of such comparisons explicit; we see no application of it that would harm subjects or communities. The main risk we can identify is the opposite of the paper's intent — that labelling claims "undetermined" could be read as licence to discount evaluation results generally — which is why §4.5 states that the observed evidence was found to be silent rather than misleading, why the paper reports *zero* refuted orderings on the real audit, and why it is explicit that the labels are conservative: an undetermined claim is one the observed evaluations do not settle, not one they contradict. On the constructive side the paper states its own limits: the acquisition rule of Appendix J is certifiably blind to the values it buys and is exact for one step, but it is myopic, it is measured on nine designed instances rather than a sample, most of its measured advantage comes from its tie-break rather than from the guarantee it is named for (§J), and it is reported in its prior-free form — G then C — rather than with the expectation-based secondary key that scores higher on these runs and needs a declared prior.

REFERENCES

Ben Aoki-Sherwood, Catherine Bregou, David Liben-Nowell, Kiran Tomlinson, and Thomas Zeng. When the universe is too big: Bounding consideration probabilities for Plackett-Luce rankings. 2024. arXiv:2401.11016.

Ofir Arviv, Kristjan Greenewald, Yotam Perlitz, Hadar Mulian, Michal Shmueli-Scheuer, and Leshem Choshen. Stop guessing when to stop testing: Efficient model evaluation with just enough data. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. arXiv:2607.08522.

Haris Aziz, Markus Brill, Felix Fischer, Paul Harrenstein, and Jérôme Lang. Possible and necessary winners of partial tournaments. *Journal of Artificial Intelligence Research*, 2015. doi: 10.1613/jair.4856.

Federico Crippa and Danil Fedchenko. Partially identified rankings from pairwise interactions. 2024. arXiv:2410.18272.

Polina Gordienko, Christoph Jansen, Julian Rodemann, and Georg Schollmeyer. Beyond Arrow: From impossibility to possibilities in multi-criteria benchmarking. 2026a. arXiv:2602.07593.

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

Anas Himmi, Ekhine Irurozki, Nathan Noiry, Stephan Clémençon, and Pierre Colombo. Towards more robust NLP system evaluation: Handling missing scores in benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. arXiv:2305.10284.

Wei-Jung Huang. What does an LLM-agent leaderboard rank actually compare? 2026. arXiv:2609.07785. IEEE DSAA 2026.

Aviram Imber and Benny Kimelfeld. Computing the extremal possible ranks with incomplete preferences. 2020a. arXiv:2005.08962.

Aviram Imber and Benny Kimelfeld. The complexity of determining the necessary and possible top-k winners in partial voting profiles. 2020b. arXiv:2002.09212.

Kathrin Konczak and Jérôme Lang. Voting procedures with incomplete preferences. In *Proceedings of the IJCAI Workshop on Advances in Preference Handling*, 2005.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35581–35602, 2025. arXiv:2410.05952.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2022. arXiv:2211.09110.

Tyler Lu and Craig Boutilier. Preference elicitation and robust winner determination for single- and multi-winner social choice. *Artificial Intelligence*, 279:103203, 2020. doi: 10.1016/j.artint.2019.103203.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Magne Mogstad, Joseph P. Romano, Azeem M. Shaikh, and Daniel Wilhelm. Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *The Review of Economic Studies*, 91(1):476–518, 2024.

Akash Kumar Mohankumar and Mitesh M. Khapra. Active evaluation: Efficient NLG evaluation with few pairwise comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. arXiv:2203.06063.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. 2026a. arXiv:2607.16259.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. 2026b. arXiv:2606.08679.

Maria Silvia Pini, Francesca Rossi, Kristen Brent Venable, and Toby Walsh. Incompleteness and incomparability in preference aggregation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1469–1474, 2007. URL `https://www.ijcai.org/Proceedings/07/Papers/236.pdf`. Conference version; the journal version is the Artif. Intell. article by the same authors.

Maria Silvia Pini, Francesca Rossi, Kristen Brent Venable, and Toby Walsh. Incompleteness and incomparability in preference aggregation: Complexity results. *Artificial Intelligence*, 175(7-8), 2011. doi: 10.1016/j.artint.2010.11.009.

Nisarg Shah and Ziqi Yu. The art of calling the winner by asking just enough questions: Competitive preference elicitation with next-best queries. 2026. arXiv:2608.29087.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195, 2010.

Jason Z. Wang. The evaluation blind spot: A stereological theory of benchmark coverage for large language models. 2026. arXiv:2606.05169.

Xumei Xi, Christina Lee Yu, and Yudong Chen. Entry-specific bounds for low-rank matrix completion under highly non-uniform sampling. 2024. arXiv:2403.00184.

Lirong Xia and Vincent Conitzer. Determining possible and necessary winners given partial orders. *Journal of Artificial Intelligence Research*, 2011. doi: 10.1613/jair.3186.

Yuchen Zeng and Dimitris Papailiopoulos. You don't need to run every eval. 2026. arXiv:2606.24020.

A PROOFS

Proposition 1. Both sides are $\forall\forall$ statements, and a universal quantifier commutes with a conjunction: i leading in every completion means for all S' and all j, $\rho(S')[i] \geq \rho(S')[j]$, which is the same as for all j, for all S'. Each inner statement is decided exactly by equation 2. Leaderhood is read *weakly* in that statement — a completion that ties i with a rival is a completion in which i leads — so it is the strict counterpart, an infimum strictly above zero, that §4.1 requires of an *established* ordering. □

Proposition 3. Write $g_{ij} = \sup_{S' \in F}\big(\rho(S')[i] - \rho(S')[j]\big)$, which is finite and, by equation 2, a sum of $|B|$ one-column closed forms; let $\mathrm{marg}(i) = \min_j g_{ij}$ over the finitely many rivals, and suppose $\mathrm{marg}(i) < 0$. Pick a rival j^* attaining the minimum, so $g_{ij^*} < 0$. No attainment argument is needed here, and none is available: the rule of equation 2 gives half credit at an exact tie and different credit immediately on either side, so ρ is piecewise constant in each cell and *discontinuous* where a cell crosses another value, and compactness of F therefore does not deliver a maximiser. The conclusion follows from the definition of the supremum instead: for *every* $S' \in F$, $\rho(S')[i] - \rho(S')[j^*] \leq g_{ij^*} < 0$. Rival j^* is strictly ahead of i in every completion, so no completion has i weakly ahead of all rivals: i is not the leader in any, and in particular is not joint-possible. Because the argument uses only $\sup < 0$, it is one-sided by construction — nothing in it excludes a row with $\mathrm{marg}(i) \geq 0$ from failing the joint statement, which is the phenomenon §4.3 measures. □

Proposition 2. Necessity is immediate: a completion witnessing the joint statement witnesses every per-rival statement. Strictness needs an instance; take four models and three columns, all four models present in every column, with the three bracketed entries unobserved and admissible range the observed span of their column, here $[0, 1]$:

	col. 1	col. 2	col. 3
i	1.00	0.00	0.00
j_1	0.00	$[0, 1]$	1.00
j_2	$[0, 1]$	1.00	$[0, 1]$
j_3	0.50	0.00	0.50

Let (x, y, z) be the free entries of j_2 in column 1, j_1 in column 2 and j_2 in column 3. The per-rival relaxation holds, and holds strictly: at $(x, y, z) = (1/2, 0, 0)$ the declared rule gives $\rho(S')[i] = 1/2$ against $\rho(S')[j_1] = 4/9$; at $(0, 1, 0)$ it gives $\rho(S')[i] = 4/9$ against $\rho(S')[j_2] = 7/18$; and at $(1/2, 0, 1)$ it gives $\rho(S')[i] = 4/9$ against $\rho(S')[j_3] = 7/18$. No completion does all three at once. Each column holds at most one free entry, and between consecutive observed entries of its column the aggregate is linear in that entry while at a tie it takes the midpoint of the two adjacent linear values, so a maximiser of $\min_j \big(\rho(S')[i] - \rho(S')[j]\big)$ exists whose free entries lie among the observed entries and the endpoints of their ranges — here $\{0, 1/2, 1\}^3$. Enumerating those 27 completions attains the maximum $-1/18$ at $(x, y, z) = (0, 0, 0)$, where $\big(\rho(S')[i], \rho(S')[j_1], \rho(S')[j_2], \rho(S')[j_3]\big) = (1/2, 1/2, 4/9, 5/9)$: i ties j_1 and trails j_3. The exact solve of §3.3 returns the same value at the same completion. (§4.3 exhibits thirteen such rows at larger scale on real measurements, each certified by that solve rather than by hand.) □

B VALIDATION OF THE CLOSED FORMS AND THE MILP

The closed-form pairwise bound of §3.3 is checked in `tests/test_identification.py` against an independent brute-force enumeration of F on instances small enough to enumerate. The enumerated extremities must lie inside the analytic ones; the closed-form column extreme must attain the enumerated optimum; and the analytic values must match hand computation on a case small enough to do by hand.

The candidate set, against exact region enumeration. The per-cell candidate set of equation 2 deserves its own check, because its soundness is not obvious: a review of an earlier draft observed, correctly, that a search over *breakpoints and endpoints alone* is unsound — at $v_i = 0.7$, $v_j = 0.3$, $[\ell, u] = [0, 1]$ it returns $\frac{1}{2}$ where the true maximum is 1 — and inferred that the implementation was unsound with it. It is not: the midpoint and mixed-midpoint terms above put a representative strictly inside the open plateau $(\min(v_i, v_j), \max(v_i, v_j))$, where the contribution is constant and maximal. Because sampling cannot establish soundness, the check is exact rather than random: the contribution of a free cell is piecewise constant with breakpoints exactly at v_i and v_j, so its extreme over $[\ell, u]$ is attained at one representative of each of the five regions intersected with $[\ell, u]$, and that representative set can be constructed directly. On 4.6×10^5 adversarial instances — both breakpoints inside, one inside and one outside, both outside, degenerate $v_i = v_j$, $[\ell, u]$ strictly inside the plateau, and both orientations of each — the implemented set matches the exact enumeration in every case, with a worst discrepancy of 0.0 (exact zero, not a tolerance). The implementation is sound; the sentence describing it in an earlier draft was not, and now states the candidate set explicitly.

What exhaustive enumeration does and does not cover. The MILP of §3.3 is checked in `tests/test_joint.py` against exhaustive enumeration of F: on those instances the possible-leader set, the always-leader set, and top-k membership for $k = 1, 2, 3$ all agree with the enumeration, as does the pruned variant of the solve that the larger instances use, and every rebuilt completion re-scores to the aggregate the solver claimed for it.

Two failure modes motivate the certification discipline of §3.3. The solver reports values with a feasibility tolerance, and a completion satisfying the integer order relation only up to that tolerance re-scores into a different ranking: on a 32-model instance we found 3418 exactly-tied pairs, and reading each tie back as a strict comparison flips half a quantile — one of $O(|M|)$ quanta, enough to move a reported comparison the wrong way. And a time-limited run may return an incumbent that is not optimal. Rebuilding from the solver's own integral relation and re-scoring under the declared rule removes both.

The sweep's own grid-search cross-check does not engage on the grid of §4.3: it fires only on instances with at most 7 free cells, and the rows there have between 17 and 109. Exhaustive enumeration therefore backs the solver's correctness on small instances, not the thirteen rows the paper reports; those rest on the solver returning `optimal` with a negative joint value, on the absence of any certified witness, and on the one-sidedness of Proposition 3, which makes an opposite error impossible rather than merely unobserved.

The thirteen negatives under a second solver, with no tolerance anywhere. The negative verdicts are the load-bearing claim of §4.3 — they are what turns the relaxation from possibly loose into demonstrably loose on published evidence — and every one of them is one solver's optimality claim about floating-point arithmetic. They are therefore re-solved by a second solver, CP-SAT, on a model written from the definition rather than translated from the MILP: cells are integers scaled by 10^8, and each ordering indicator is *defined* by a reified comparison ($z = 1 \iff u_k \geq u_j + 1$ grid step, $z = 0 \iff u_k \leq u_j$) in place of the big-M rows, so neither a tolerance nor the separation constant δ appears in it at all. It reproduces all thirteen values to machine precision, the largest disagreement with the stored artifact being 3×10^{-16} win-rate units, and returns the same sign on each. Twenty-four rows the artifact records as possible-leader are re-solved the same way as controls, since a check that could only ever return *negative* would agree with thirteen negatives for the wrong reason: they agree to 7×10^{-16} and not one comes back negative. The re-solve is *less* restrictive than the paper's model rather than more — it admits every completion in the ranges, where equation 3–equation 5 admit the δ-separated ones — so a negative there implies the negative reported here, and the resolution is not a hidden restriction: the objective depends on the per-column ordering pattern alone, observed values and range endpoints are exact at 10^{-5}, and the smallest positive gap between distinct observed values is 10^{-3}, five orders of magnitude above the grid step. `paper/_gapindependent.py` runs the comparison. OR-Tools is a verification-only dependency: nothing in `tools/` or `experiments/` imports it and no number the paper reports is produced with it.

C PUBLISHED PRECISION AND THE RULE REPLICATION

The published matrix loses no information. HELM's public release group tables publish the same quantities as full-precision floats. Matching on model and scenario name, published $=$ $\mathrm{round}(\text{release}, 3)$ holds on *all* 876 published cells of the `classic: core_scenarios` Accuracy universe and *all* 704 cells of the `classic: targeted_evaluations` block of §4.2, with zero mismatches, zero published cells lacking a release value, and zero release values lacking a published cell. The published matrix therefore contains no information the release lacks, and every identification result in this paper is exact conditional on two pinned reconstructions — the recovered matrix and the reimplemented rule — rather than on the historical published aggregate, which the pinned rule reproduces to 0.019 and not exactly (§4.1).

One scale detail is worth recording because getting it wrong corrupts exactly three columns. The published matrix stores Accuracy in the percent units its columns declare, so a fraction-valued metric is rescaled by 100 (MMLU $0.5089123 \to 50.891$). Not every Accuracy column is fraction-valued: a bits-per-byte metric is a per-byte entropy and is published at its native scale (The Pile $0.5925242 \to 0.593$, ratio 1.0). Applying the rescale unconditionally would corrupt exactly the three bits-per-byte columns. The audit does not assume the scale: it is estimated per column from the ratio to the published values and asserted to be 1 or 100, so a column following neither convention stops the audit rather than passing through it.

The residual is not a precision artefact. Re-run on the recovered full-precision values, the pinned rule still differs from HELM's published aggregate column by at most 0.019 in win-rate units (median 0.0017; 4408 of 4422 pairwise comparisons concordant; identical top-10 ordering). On the rounded matrix the same statistics are 0.019 maximum, 0.0017 median and $4410/4422$. The discrepancy is therefore not caused by published precision, and we report it as an unresolved difference between the pinned implementation and the platform's published column.

Variants of the rule tried against the same published column, with the pinned rule ($n_b - 1$ denominator, ties at half credit) as the reference (0.0189 max, 0.0017 median):

Variant	max |residual|	median
ties scored strictly (credit 0)	0.0189	0.0015
ties scored fully (credit 1)	0.0157	0.0106
denominator n_b (drop the -1)	0.0220	0.0074
drop columns with $n_b < 20$	0.0189	0.0017
drop columns with $n_b < 32$	0.0459	0.0027
drop columns with $n_b < 40$	0.0600	0.0041
median rather than mean aggregation	0.1906	0.0265
input rounded to $3, 4, 5, 6$ decimals	0.0189	0.0017

No variant attains a smaller maximum residual than the pinned rule; one (strict tie credit) attains the same maximum and a marginally smaller median, and it is reported here rather than adopted, since adopting it would change every win rate in the paper on the strength of a 0.0002 difference in one statistic. Phantom rows and the remaining column-coverage cutoffs were tried and are no better.

Provenance of the rule, and of the residual. The rule is not a reconstruction from documentation. `summarize.py::compute_aggregate_row_win_rates` was retrieved from the project's public repository, and the implementation in `tools/identification.py` matches it line for line: the columns it skips because their direction is null, the columns it skips because fewer than two values are present, the tie credit $\mathrm{acc}+(c-1)/2$ accumulated in sorted order, the n_b-1 denominator, and the unweighted mean over the columns a row is present in. The release table's own header was checked against one consequence of that source: all sixteen Accuracy columns declare `lower_is_better` explicitly false, so the null-direction skip is not engaged here and cannot be the source of the gap. Two further candidate explanations were tested and eliminated rather than left open. It is not the row set: recomputing the pinned rule on the release table's own 67 rows, with no model restriction at all, reproduces none of the 67 published values exactly and leaves the same 0.019 maximum and the same $4408/4422$ concordance. And it is not input precision: re-running on inputs rounded to 3, 4, 5 and 6 decimals leaves the residual unchanged. What remains is an unresolved difference between the documented rule and the column the platform published, and we report it as such.

What the residual can and cannot do to the labels. The discrepancy is left unresolved — we have not established which module of the release emits the published column — so the honest question is not whether it exists but whether it can move a label. It cannot, and this is checkable rather than argued. The residual shows up in exactly 14 *ordered* comparisons, which are the two orientations of seven unordered pairs: (Falcon (7B), RedPajama-INCITE-Base (7B)), (J1-Grande v2 beta (17B), Vicuna v1.3 (13B)), (T0pp (11B), Pythia (6.9B)), (Cohere medium v20220720 (6.1B), text-babbage-001), (TNLG v2 (6.7B), RedPajama-INCITE-Base-v1 (3B)), (ada (350M), text-ada-001) and (MPT (30B), MPT-Instruct (30B)). *All fourteen are undetermined at all three assumption levels*, and the narrowest interval among them — ada against text-ada-001 at A2, $[-0.0824, +0.0104]$, a width of 0.093 — is nearly five times the maximum residual of 0.019; the widest is 1.50. So no disagreement between the pinned rule and the published column falls on a comparison whose label is decided, and the reported top-ten order is the same under both. The labels of Table 1 are therefore exact for the pinned reconstruction and *unchanged* if the published column is taken as the target instead: the two audits differ in the rule they audit, not in what they conclude about the reported order.

One count in this paper does depend on the pinning, and we report it rather than bury it: applying the pinned rule to the *published* three-decimal matrix rather than to the recovered full-precision values raises the established count from 4 to 5 at A0 and A1 (A2 is unchanged at 6), because a single thin ordering — `text-davinci-003` over `TNLG v2 (530B)`, whose lower bound is 0.000473 at full precision — crosses the zero margin once the inputs are rounded. §4.1 quotes the full-precision figure, which is the more conservative of the two, and the published-matrix arm is stored beside it (`pilot_04_precision/results.json`, keys `headline_published` and `headline_full_precision`).

Metric direction. Those same three columns are also the only ones whose *direction* the two sources disagree about: HELM's leaderboard marks the bits-per-byte metrics (The Pile, TwitterAAE, ICE) `lower_is_better` while the analysed matrix marks all 207 of its columns `higher_is_better`. Appendix F resolves that conflict in HELM's favour and quantifies what the

choice moves. The `classic: core_scenarios` Accuracy universe contains none of the three, so §4.1 is unaffected either way.

D A SECOND AND THIRD DECLARED UNIVERSE

§4.5 fixes the empirical scope of this paper: one platform, one aggregation operator, one historical release. Two parts of that scope are not removable from the data in hand — a rule with genuine cross-column coupling, an Elo update say, is outside the proved rule class, and no second platform's matrix is present here — but the *declared benchmark universe* is not fixed by the same constraint, and treating it as fixed would have overstated the limitation. HELM's own release group table for `classic: core_scenarios` declares several universes side by side, each carrying its own aggregate column produced by the *same* mean-win-rate operator over a different declared set of measured quantities. Three of them are in this paper's data: Accuracy, which §4.1 audits, plus Robustness and Fairness. The second instantiation was therefore available at no new data cost, and has been run (`pilot_05_universes/results.json`) rather than argued to be unavailable.

declared universe	scenarios	models	observed	established / 45			possible	identified
				A0	A1	A2	A1 (relax.)	any level
Accuracy (audit of §4.1)	16	67	81.7%	4	4	6	16	0
Robustness	14	67	84.4%	0	0	13	11	0
Fairness	14	67	84.4%	3	3	12	9	0

The "possible" column is the *cheap per-rival relaxation* — an upper bound that needs no solver and certifies nothing — not the exact joint solve, and it is labelled as such here because the distinction is the one the audit of §4.1 turns on. Two of the three rows carry no certified count at all: the joint solve behind the certified column of Table 1 was run on Accuracy and not re-run on Robustness or Fairness, so for those two universes the paper reports an admitted count and no certification, rather than presenting an upper bound as a finding. On Accuracy the same relaxation admits 16 at A1 against 14 certified possible and 2 undecided, and the gap is exactly the undecided set.

The counts are of the 45 pairings inside each universe's reported top ten, so the established fraction is $4/45$, $0/45$ and $3/45$ at A1 — the same order of magnitude, not a coincidence of one slice. In all three universes the reported leader is the same model, `Llama 2 (70B)`, and in none of the three is it identified: its worst-case margin against its best rival is -0.348 (Accuracy), -0.234 (Robustness) and -0.256 (Fairness) at A1, so every completion of the declared universe contains a rival that beats it. Nothing here is established about the *leader* in any universe, and the admitted-leader count is the only leader-side statement the evidence supports here.

Both new universes are audited on the paper's own precision arm. Reporting a second universe computed on HELM's published three-decimal matrix alongside a first computed on the recovered full-precision values would mix arms, and the arms are not interchangeable: on Accuracy they disagree, giving 5 established orderings at A0 and A1 instead of 4 (§4.1, Appendix C). The recovery gate is therefore applied to each new universe exactly as §4.1 applies it to Accuracy. For both, every published cell has a release value and published $= \mathrm{round}(\text{release}, 3)$ holds on all 792 cells with zero mismatches, so the published matrix loses no information there either. The pinned rule's residual against HELM's own published aggregate is a maximum of 0.0455 (Robustness) and 0.0405 (Fairness) in win-rate units, medians 0.0014, with identical top-ten orderings in both — the same unresolved, non-precision residual as Accuracy's 0.019, of the same order and with the same character. And unlike Accuracy, the two arms *agree* in these universes: every headline count in the table above is identical whether it is computed on the recovered values or on the published matrix, at all three levels. The smallest |infimum| among the reported top-ten pairs is 0.0049 in Fairness and 0.0054 in Robustness, which is ten and eleven times the half-quantum of 0.000473; no reported pairing sits within half a win-rate quantum of flipping, which is the sense in which the two arms agree here rather than merely happening to round to the same integers.

What is still not available. This appendix buys a second and third declared benchmark universe under the same platform and the same operator. It does not buy a second platform or a second operator, and the paper does not claim one. What carries across platforms is Proposition 1 and

Level	established	refuted	undetermined	possible ($\geq$)	undecided	worst-case margin
A0	4	0	41	12	8	-0.354
A1	4	0	41	14	2	-0.348
A2	6	0	39	8	1	-0.308

Table 1: Labels for the 45 pairwise orderings within the reported top ten of HELM's declared `classic: core_scenarios` Accuracy universe under three assumptions about unobserved scores, on the recovered full-precision matrix (Appendix C). "Possible" counts models the exact joint solve *certified* able to lead in some completion, a lower bound since "undecided" rows are ones the solver left time-limited, where the cheap per-rival relaxation would have admitted 20, 16 and 9. Identified leaders are 0 at every level (Proposition 1); the worst-case margin is $\inf_{S' \in F}$ of the reported leader's aggregate minus its strongest rival's.

equation 2, which hold for any rule that is a column-wise mean of a function of that column alone; what is demonstrated is HELM's mean win rate on three of its declared universes. A platform whose aggregate couples columns — the case §4.5 names — is outside both the demonstration and the proof, and would need the joint solve rather than the closed form.

E THE AUDIT'S LABELS, AND THE COMMON-SUPPORT CONTRAST

Table 1 records the labels behind §4.1 at each assumption level, on the recovered full-precision matrix; the figure below draws the same labels as bars, beside the leader question and the complete-case remedy. Because the contribution is claim-level — *which* comparisons the evidence settles, not how many — Table 2 lists all 45 reported top-ten orderings individually, with the infimum of the aggregate difference at each level and that level's label, so the counts of Table 1 can be read off one comparison at a time.

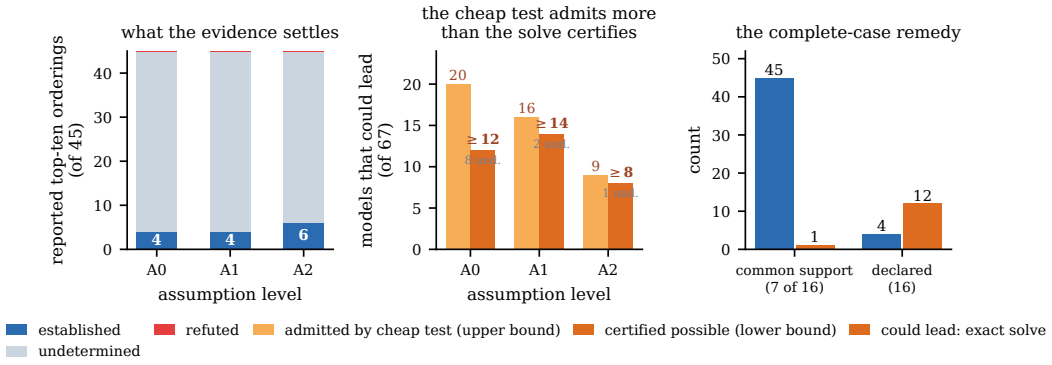

Figure 3: Each panel answers one question at the three assumption levels, read from the artifacts behind Table 1. Left: the labels of that table as bars, counted against the 45 pairwise orderings inside the reported top ten; no ordering is refuted at any level, so the red segment is absent. Middle: how many of the 67 declared models could lead in some completion, in three separately counted categories that are never summed here. *Certified possible* (the exact joint solve returned a re-scored witness) is 12, 14 and 8 at A0, A1 and A2. *Undetermined* (the solve hit its time limit before a proof, which is not a negative result) is 8, 2 and 1. *Admitted by the cheap per-rival relaxation* — an upper bound that needs no solver and certifies nothing — is 20, 16 and 9. The third count exceeds the first by exactly the second at every level, and the same models: every row the relaxation admits and the solve did not certify is one the solve left undecided, and on this universe the solve never refuted the relaxation. Right: the complete-case remedy, which establishes all 45 orderings and leaves a single possible leader over the 7 of 16 scenarios every declared model ran, against 4 of 45 and 12 certified-possible leaders under the declared universe at A0. The right panel is not a second view of the same comparison: the restriction discards 9 of 16 scenarios and 46% of the observed cells.

The reported order interleaves coverage sets. The reported leader `Llama 2 (70B)` is averaged over *10* of the 16 scenarios at mean win rate 0.9436, while three of the five models immediately

below it (`text-davinci-002`, `Cohere Command beta (52.4B)`, `text-davinci-003`) are averaged over all 16 — the reported order therefore interleaves models scored on different benchmark sets, each normalised against a different population of comparisons.

The two deletion remedies, measured on the same currency. §1 contrasts the paper's route with deletion and §4.1 prices the common-support restriction. Both figures come from the same declared universe and the same counts, so they are comparable, and Table 3 sets them side by side. A completely measured block makes the feasible set a singleton, so on each restriction the rule is evaluated once and no bound is needed; "possible" and "identified" therefore coincide there, and neither remedy is doing identification work. What separates them is which models and which scenarios they throw away. The model-dropping row is read from `pilot_07_deletion/results.json`; the column-dropping row and the reported universe are the common-support and audit runs of §4.1.

F METRIC DIRECTION: WHICH CONVENTION IS AUTHORITATIVE

The analysed matrix records all 207 of its columns `higher_is_better` over a declared $[0, 100]$ percent range. HELM's own leaderboard — the artefact this paper studies — declares *exactly three* of those columns `lower_is_better`: the bits-per-byte metrics The Pile, TwitterAAE and ICE, all in the `classic: targeted_evaluations` Accuracy family. The two declarations cannot both describe the metric, and what decides between them is the metric itself. A bits-per-byte score is a compression loss, so lower is better by definition; and the three columns' values here, 0.466 to 3.505, are per-byte entropies rather than percentages, so the blanket `pct` / $[0, 100]$ record is already known to misdescribe them on scale alone (Appendix C). **We therefore treat HELM's declaration as authoritative**, and the masking study is reported under it. Following the matrix's record instead would rank 3 of that block's 22 scenarios by an ordering the benchmark does not use, and would let a metadata defect decide the study's own ground truth — the true order on the fully measured block.

Which slices this affects. The three columns are not spread evenly across the paper's two real-data settings. The `classic: core_scenarios` Accuracy universe of §4.1 contains *none* of them, so every audit number in this paper is invariant to the choice and needed no re-analysis. The `classic: targeted_evaluations` Accuracy block of §4.2 contains all three of its 22 scenarios, so the masking study is where the choice binds.

What the choice moves. On the unmasked 32×22 block the reported leader is the same model under both conventions (`text-davinci-003`), but the ordering is not: the top-ten *set* changes (`OPT (175B)` enters it at rank 9 where the matrix's convention ranks it 14th, and `Cohere xlarge v20221108 (52.4B)` leaves it), the largest per-model win-rate shift is 0.136, and the largest rank shift is 8 (with 11 of the 32 models moving by 3 ranks or more). Over the 280 masked analyses the established count is higher under HELM's convention (mean 2.91 against 2.05 at A1; largest single-setting difference 6.1 of the 45 orderings). Rows of the two arms are not comparable one by one, because the two conventions rank a different set of models into the true top ten, so the 45 orderings being labelled are not the same 45. Three conclusions are invariant and hold under *either* convention: zero true top-ten orderings are reversed, zero models are identified in every completion, and the true leader is a possible leader in all 280 instances. The native rule's leader is the true leader in 188 of the 280 instances under HELM's convention and in 183 under the matrix's. Both arms are stored side by side, the authoritative one under the plain name (`exp_b_masking/results.json`) and the matrix's convention as the labelled sensitivity arm (`results_published_convention.json`).

Both arms are one computation, at the level the study reports. The two conventions are run by a transformation that is exact rather than approximate: negating a lower-is-better column's values reproduces its direction flag exactly, ties included, because negation is a bijection on the column's values that reverses the order in which upstream visits them — which is precisely what the flag does. The transformation acts on the observed values, so it maps A1 to A1 exactly: the range A1 builds from a column's kept values is the negated range of the other arm's, and the arms are one computation rather than two. The equivalence is asserted in `tests/test_identification.py`.

A0, by contrast, is not available for this block, and negating the matrix's recorded range would not have made it available. A0 is the range the platform *declares* for a metric, and the only declaration

on offer for these three columns — a percent range — is the one this appendix has just shown not to describe them, on scale as well as on direction. Negating a range whose justification has been withdrawn does not supply one: it would assert that the loss lies in $[0, 100]$ bits per byte, a bound no source states, and a bound that no other choice of endpoint would make less arbitrary either, since the metric's domain has no finite maximum that anyone declares. §4.2 therefore reports A1, which needs no declaration at all. The $A0$ records remain in both artifacts, where they are the matrix's own percent record (negated for the three columns, in the arm that follows HELM's direction), kept for reference; no number in this paper is taken from them.

G THE MASKING STUDY: CALIBRATION DETAIL

Table 4 is the grid the main text summarizes, and this appendix records what the calibration could and could not match. Every number in it is at A1: three of this block's scenarios are bits-per-byte metrics whose recorded percent range does not describe them, so the declared range A0 needs does not exist for the block and no A0 is reported for it (Appendix F).

The three structured patterns saturate at the dense end of the coverage range. Block cannot exceed 74.5% realized coverage: its scale is set by the real block-size distribution, whose groups share fixed column sets, so no setting of the scale parameter produces a denser instance, and the two settings aimed at 90% and 80% coverage land on the same instance at 74.5%. Budget and column cannot reach their 90% targets either (79.1% and 83.3%: the drawn per-model and per-column coverage distributions clip at one), and budget misses its 80% target by 2.1 points. A setting counts as met when realized coverage lands within 0.4 points of its target — the stopping rule the mask search itself uses (`exp_b_masking/exp_b.py`) — and 19 of the 28 meet it. The other nine miss: seven by more than a point (the worst by 15.5 points) and two by less than one.

The consequence is a limit on which comparisons are matched. The band in which all four patterns have settings close to a common coverage is roughly 30%–70%, and the two contrasts the main text draws sit at or just above its top: uniform deletion at 79.9% against budget at 77.9%, matched to within 2 points; and block at 74.5% against uniform deletion at 70.1% and at 79.9%, which *brackets* block's coverage rather than matching it — the bracketing is what that comparison needs, since block there retains more established orderings than uniform deletion manages at a coverage below its own as well as at one above it. We have not extrapolated any pattern outside the coverage range its settings occupy.

How often A1's feasible set contains the truth. A1's range for a column is built from that column's *observed* cells, so a hidden score outside the observed span is excluded from F: this is the defect §3.2 names, and the masking study is the one place it can be measured, because there the hidden scores are known. Across the 280 instances 14.5% of the $81{,}884$ hidden cells fall outside their own column's observed span — 8.7% at the lowest missingness rate and 21.3% at the highest — so the true completion lies inside F in 1 instance of the 280, and in 277 of the remaining 279 the exclusion reaches a true-top-ten row. The zero reversal count of §4.2 is therefore a measurement rather than a consequence of the feasible set admitting the truth: A1 is unsound in principle in nearly every instance, and no true ordering is reversed all the same. `paper/_maskfeas.py` recomputes each count from the study's own masks.

Seed dispersion, and two rival explanations. Every cell of the table is a mean over ten seeds, so the question is whether a contrast is a difference of means or an ordering that holds draw by draw. It is answerable exactly rather than statistically: the pairwise-exact analysis is deterministic given a mask and the mask is deterministic given a seed, so the only randomness in a cell is which seed, and there are ten of them. A second run (`exp_b_dispersion`) recomputes the four settings the contrasts use — the three named above plus the block pattern — seed by seed, and asserts that each setting's ten records average to the number already in the table, so a spread is never published for an instance the table does not describe.

The main contrast survives: at the same missing rate the budget pattern establishes strictly more true top-ten orderings than uniform deletion in *ten seeds of ten* (13.1 against 2.4 on average), and it does so at *lower* realized coverage (77.9% against 79.9%). The bracketing contrast survives on the side the main text uses and only on the mean on the other. Block against uniform deletion at 70.1%

is eight seeds strictly above and two ties, both at zero; block against uniform deletion at 79.9% is seven above, one tie and *two below*, so the mean advantage there (6.2 against 2.4) is a mean and not an ordering that holds in each draw. The bracket is what the comparison needs, and the lower side is the side doing the work.

The coarsest rival explanation is that structured patterns simply leave more complete cases, and a complete case is a comparison that needs no bound at all. For complete *columns* that is void rather than merely unsupported: at 70–80% coverage of a 32×22 block not one of the forty draws leaves a single fully observed column, so no ordering established anywhere in this study rests on one. Complete *models* are real and differ by pattern — the budget pattern leaves 13–22 fully observed models, the block pattern 1–9, uniform deletion 0–1 — so their count does track the ordering of the means. The seeds separate it from being the explanation: one block draw establishes 11 orderings from 3 complete models while one budget draw establishes 4 from 13. What the block pattern leaves instead is shared structure, and it is what the mechanism above is about: its largest group of models with an identical evaluated-scenario set holds six models (9 in one draw), and models inside such a group are mutually comparable with no complete model anywhere.

Matched coverage, and what the pattern ordering rests on. The table compares the patterns at matched *missing rate*, and `make_mask` calibrates each pattern's scale so that its realized coverage lands near the target — but the calibration is not exact, and it does not need to be, except that it leaves cross-pattern sentences comparing cells whose coverages differ: at the 10% rate the block pattern reaches 74.5% and uniform deletion 89.8%. A reader is therefore entitled to ask whether the ordering describes the *pattern* or only the amount of data. A third run (`exp_b_dispersion/matched_coverage.py`) answers it by rebuilding all 28×10 masks, asserting their realized coverages against the ones already in the table, and reading each pattern's established count as a function of coverage by linear interpolation between the measured grid points.

On the window the four patterns share, and restricted further to the sub-window in which any of them establishes anything, the ordering is the same at every level surveyed:

coverage	budget	block	column	uniform
70.0%	5.33	2.33	0.11	0.00
72.0%	7.29	4.05	0.85	0.47
74.0%	9.24	5.78	1.60	0.96

So *pattern matters beyond amount* survives, and in a sharper form than the table alone shows. The block-against-uniform sentence of §4.2 is quoted across a 4.4-point coverage gap; at matched coverage the block pattern's 6.20 at 74.5% stands against an interpolated 1.08 for uniform deletion rather than the 0.00 it reaches at 70.1%. What does *not* survive is the stronger reading in which the block pattern is the most concentrated of the four: at every matched level the per-model budget pattern is ahead of it. Two caveats belong with the table. Every level in it is interpolated for at least three of the four patterns — 74.5% is a measured point only for the block pattern — so the interpolation is a reading of the measured grid and not a measurement; the same artifact reports the directly measured cross-pattern pairs with their coverage gaps. And interpolation can only restate the grid's shape, so what the table establishes is that the ordering does not reverse over the window, not what the curves do between the grid points.

The complete-case rival explanation behaves the way the seed-level analysis found at matched coverage too: the number of fully observed models is concordant with the established ordering in 15 of the 18 informative pairwise comparisons at these levels, discordant in 2 and tied in 1; both discordant pairs are uniform deletion against the column pattern, where the column pattern establishes more from no more complete models. The coarsest version of the explanation is refuted by a measurement that needs no interpolation at all: at the 10% rate the column pattern leaves 15 whole columns fully observed — every comparison inside them needing no bound whatsoever — and establishes 9.7 orderings, while uniform deletion at the 10% rate leaves 2 complete columns and establishes 20.0.

The level choice does not move the conclusions. The A0 records this study does not report are kept in the artifact, and they answer the obvious objection to reporting A1: across the 28 settings, moving from A0 to A1 changes the established count by at most 1.9 of 45 (median 0.0), and the

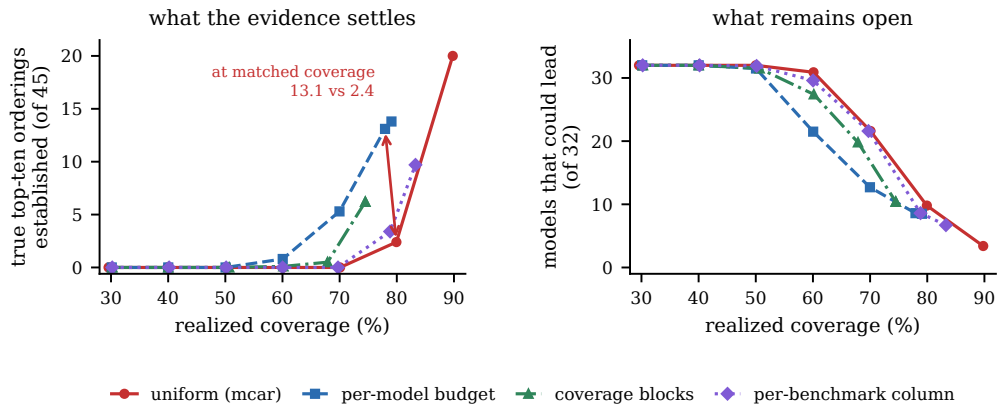

Figure 4: Identification against realized coverage under four hiding patterns, at A1, means over 10 seeds. The further right a point sits, the more complete the matrix. Left: how many of the 45 true top-ten orderings the surviving evidence establishes. Right: how many of the 32 models could lead in some completion. The arrow marks the central contrast drawn in §4.2 — uniform deletion at 79.9% coverage against the per-model budget pattern at 77.9%, two points *less* complete and more than five times more decisive.

identified count is 0 at both levels. The asymmetry §4.1 found on the published matrix is reproduced on controlled data rather than created by the level.

H THE JOINT SOLVE: FORMULATION, TIES, AND THE NEGATIVE CERTIFICATE

This appendix gives the MILP behind every "possible leader" count and behind the thirteen negative results of §4.3, at the level of detail needed to check why an optimal negative value certifies that *no* completion realises the claim. It describes the computation that was run; the code is `tools/joint.py`.

Variables. For each column b and each model k whose cell (k, b) is unobserved, a continuous $u_{k,b} \in [\ell_b, u_b]$, the value the completion assigns it. A cell that *is* observed contributes no variable: its value is a constant. For every ordered pair (k, j) and column b with at least one free endpoint, a binary $z_{k,j,b} \in \{0, 1\}$, read as "k is strictly above j in column b". For every rival $r' \neq r$, a binary $y_{r'}$. One continuous $t \in [-1, 1]$.

The ordering relation. For every ordered pair (k, j), every column b, with $\delta = 10^{-4}$ and $M = 2$:

$$u_{k,b} - u_{j,b} \geq \delta - M\,(1 - z_{k,j,b}) \tag{3}$$

$$u_{k,b} - u_{j,b} \leq M\, z_{k,j,b} \tag{4}$$

$$z_{k,j,b} + z_{j,k,b} \leq 1. \tag{5}$$

Both directions are required. With equation 3 alone, $z_{j,k,b}$ is unconstrained whenever $u_{k,b} < u_{j,b}$ and the solver may set it to whatever the objective prefers, silently detaching the linearised aggregate from the ordering it is supposed to describe; equation 4 closes that. Ties are representable and are *exact*: $z_{k,j,b} = z_{j,k,b} = 0$ makes equation 4 read $u_{k,b} - u_{j,b} \leq 0$ in one direction and $u_{j,b} - u_{k,b} \leq 0$ in the other, which jointly force $u_{k,b} = u_{j,b}$, while equation 3 is then slack in both directions, its right-hand side being $\delta - M < 0$. δ exists only to make "strictly above" a strict inequality in a solver that works with tolerances.

What δ costs, and why it costs nothing here. Read as a condition on the completion rather than on the solver, equation 3–equation 5 represent a column's values only when they are equal or at least δ apart: a smaller strict difference fits neither the tie case, which forces equality, nor the strict case, whose margin is δ, so the encoding describes the δ-separated completions and is silent about the rest. That condition is measured rather than assumed. Over all 22 columns of the masking block and all

16 of the audit universe, the smallest positive gap between two distinct observed values is 10^{-3}, ten times δ, so a free cell that has to sit strictly between two observed values has room to do it; and the negative verdicts themselves are nowhere near the boundary, the thirteen certified negatives of §4.3 running from -0.0045 to -0.040, the closest of them forty-five times δ from zero. The separation is a representability condition on the completion, not a numerical fudge, and on this data it is met with an order of magnitude to spare.

The objective. With $W_{k,j,b} := \frac{1}{2} + \frac{1}{2} z_{k,j,b} - \frac{1}{2} z_{j,k,b}$ — which is 1 for a strict win, $\frac{1}{2}$ for a tie, 0 for a loss, exactly the win-rate contribution of equation 2's rule — the completed-universe aggregate is linear in z:

$$\mathrm{agg}(k) \;=\; \tfrac{1}{2} \;+\; \frac{1}{2B(N-1)} \sum_{b \in B} \sum_{j \neq k} \big(z_{k,j,b} - z_{j,k,b} \big),$$

because on the completed universe every column has all N models present, so every model's aggregate is a mean over the same B columns with the same denominator $N - 1$. This is the one place the joint solve differs from the pairwise bounds: the bounds of equation 2 are exact for *differences*, but a leader claim is a conjunction over rivals, and z is shared across the rivals it must satisfy.

Possible leaderhood. r is in the top-k in some completion if and only if the optimum of

$$\text{maximise} \quad t \tag{6}$$

$$\text{subject to} \quad \mathrm{agg}(r) - \mathrm{agg}(r') \;\geq\; t - M\, y_{r'} \qquad (r' \neq r) \tag{7}$$

$$\textstyle\sum_{r' \neq r} y_{r'} \;\leq\; k - 1 \tag{8}$$

satisfies $t^* \geq 0$. Setting $y_{r'} = 1$ *relaxes* equation 7 for that rival, so the cardinality row equation 8 says that at most $k - 1$ rivals may be left unconstrained — exactly the statement that at most $k - 1$ models may end up above r. t is bounded below by -1 rather than 0: clamping it at 0 would make the test $t^* \geq 0$ vacuously true and report every row as a possible leader. At $k = 1$ this is the paper's leader query, and equation 8 reduces to $\sum_{r'} y_{r'} \leq 0$, so every rival is constrained at once — which is what makes the problem joint rather than pairwise.

When a negative answer is certified. A negative verdict is reported only when HiGHS returns status 0 (optimal) *and* $t^* < 0$ at a tolerance of 10^{-7}. Any other status — in particular status 1, the time or iteration limit — yields `undecided` and no verdict, whether or not the solver returned an incumbent; an incumbent cannot certify impossibility, only possibility. This is why Table 5 reports 92 time-limited rows as unresolved rather than as non-gaps, and why the possible-leader counts of Table 1 are lower bounds. Positive answers are certified independently of the solver's arithmetic: the completion is rebuilt from the solver's own *integral* z relation and re-scored with the declared rule of equation 2, and is accepted as a witness only if re-scoring reproduces the claim.

Solver settings and the direction the tolerances err in. No tolerance is overridden: the only options passed to HiGHS are `presolve` and, where a limit is wanted, a time limit, so the solver's own defaults apply — 10^{-7} on primal and dual feasibility, 10^{-6} on MIP feasibility, and a MIP relative gap of 10^{-4}. Two settings are the paper's own. The ordering indicator equation 3 carries the strict separation $\delta = 10^{-4}$, two orders above the MIP feasibility tolerance, so a margin the solver can blur is never asked to carry a strict inequality. And the big-M constants are sized from the range they relax rather than tuned: per column $M = \max(\text{observed span},\ \text{declared width}) + 1$, which dominates $|u_k - u_j|$ by at least 1, and 2 on the aggregate rows, where the aggregates are win rates in $[0, 1]$ and $t \in [-1, 1]$, so a relaxed row holds at the bound rather than only for large M.

A negative verdict is read from the solver's own optimality claim, so the gap criterion matters, and it is safe in the direction that counts: a relative gap of 10^{-4} leaves the dual bound within that factor of the incumbent, and a negative incumbent scaled by $(1 - 10^{-4})$ is still negative, so the tolerance cannot turn an undecided row into a negative one. The read-out thresholds are asymmetric in the same direction: the negative test is $t^* < 0$ at 10^{-7} while the witness is read from the binaries at $\frac{1}{2}$, so a row whose optimum lies within 10^{-7} of zero is returned as a possible leader and never as an impossible one. An error in the other direction — a negative certificate manufactured from a row the solve did not decide — is what the status rule above prevents, and `tests/test_joint.py` exercises it directly: a time-limited solve with no incumbent must raise rather than return a negative.

The MIP feasibility tolerance leaves one measurable trace on the reported *values*, which is worth naming because the values are what the thirteen negatives are read from. The objective is linear in z and z is integral, so an exact optimum is a multiple of $1/(2B(N-1))$; of the 1,811 decided rows 1,801 are, and the remaining ten sit exactly 10^{-6} above that lattice — the tolerance itself, in the direction that makes a negative verdict *harder* to certify. All ten are possible-leader rows, whose verdict is read from a re-scored witness rather than from the objective, so no claim here rests on any of them. Appendix B re-solves the thirteen with no tolerance at all.

Exact reductions. Two reductions are applied before the solve, and both are exact rather than heuristic. A rival r' whose pairwise supremum $\sup_F(\mathrm{agg}(r') - \mathrm{agg}(r))$ is below 0 can never sit above r in any completion, so $y_{r'}$ is fixed to 0; and t is bounded above by the k-th largest pairwise supremum, which is a valid upper bound on the joint optimum and prunes the tree immediately instead of letting the solver rediscover it. Both are stated as valid only because the pairwise bounds they rest on are exact, and the pruned and unpruned solves are checked against each other and against exhaustive enumeration in `tests/test_joint.py` (Appendix B).

I THE JOINT-VERSUS-PAIRWISE GRID: WITNESSES, BANDS AND SOLVER LIMITS

Table 5 gives the accounting of the 2,400 leader queries, and Table 6 the incidence of the sign flip by level and universe size; this appendix lists the rows themselves, the coverage bands and the solver's limits. Each row is solved twice, once for $\max_{S' \in F} \min_j(\rho(S')[r] - \rho(S')[j])$ and once for a completion in which r leads, certified by re-scoring under the declared rule rather than by trusting the solver (§3.3). The instances are masked sub-universes of the 32×22 block of §4.2 under the block mask only, at A1 and A2 with declared universe sizes $n \in \{5, 6, 8, 10\}$ and $n \in \{5, 6, 8\}$; A0 was not run, so every incidence here is conditional on A1 or A2.

Why the screen runs at $k = 1$, and why the grid varies what it varies. A completion that places r first places it in the top k for every k, so the exact top-k feasible set is nested increasing in k; the relaxation $\min_j \sup_F$ does not depend on k at all, so the screen admits the same rows at every k and the admitted-but-unrealisable rows are most numerous at $k = 1$. Ruling a row out there therefore rules it out everywhere: absence of a gap at $k = 1$ implies absence at every k. The grid varies universe size and missingness rate because a gap requires free cells in *rival* rows — the monotone instrument of §4.3 — so both variables change how many free cells a rival has.

The undecided rows, in full. Screening is concentrated at *low* missingness, where a dense matrix lets the cheap test exclude rows without a solve, and budget exhaustion at the sparse end; both are most visible in the $|M| = 10$ cell at A1, whose 500 rows carry every one of the grid's 92 timeouts. There, 40 and 31 of the 100 rows of a rate setting are excluded by the screen at 20% and 30% missing against 0 at the three higher rates, while 24 rows exhaust the budget at 50% missing, 48 at 70% and 19 at the sparsest setting, the 30% setting contributing one. The other six cells account for none of the 92 (Table 6), so the undecided rows are not a random sample of the grid.

The $n = 12$ probe. On a probe configuration at 74% coverage the solver returned 9 of 12 rows in 35 seconds, but at 47% coverage 1 of 12, and at 20% coverage none of 12, the timeout-dominated configurations costing about six minutes each. The full $n = 12$ block was therefore not run.

The cost basis behind §4.1. Both costs quoted there are per row of one assumption level on the audit's 67×16 declared universe, and both are wall clock on the machine used rather than a controlled benchmark: the ratio transfers, the seconds do not. The closed-form figure is timed fresh: the screen needs the pairwise bounds of equation 2 for every ordered pair, which the production path computes once per level and reads off — two runs on an idle machine gave 109 and 133 s over all three levels, or 0.54 and 0.66 s per row-level, and a third taken while another job held a core gave 173 s, so the figure moves by roughly a factor of 1.6 with load. The joint-solve figure is the recorded per-level wall clock of `pilot_02_joint/results.json` divided by the rows that level admitted to it ($2063/20$, $1191/16$ and $784/9$ seconds), giving 74–103 s per row-level. The gap of over two orders of magnitude on the idle runs — no less than eightyfold even on the loaded one — is the

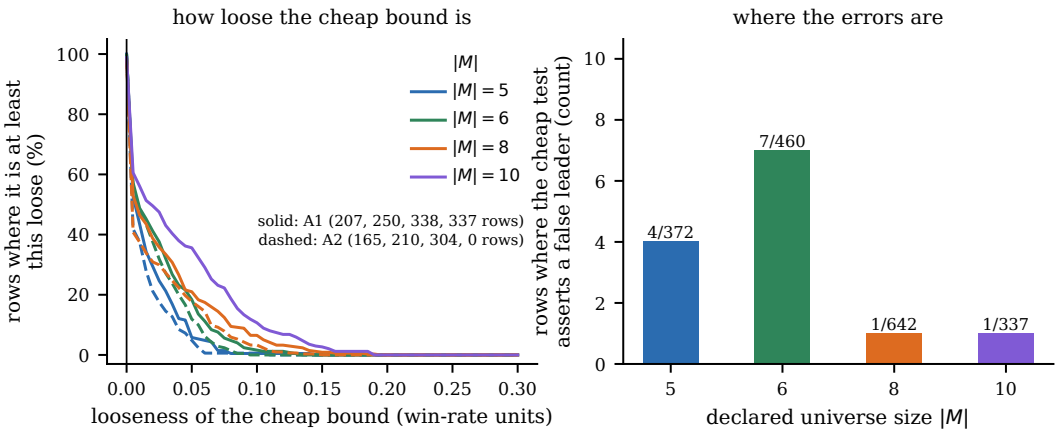

Figure 5: The cheap bound's error over the whole grid. Left: the cumulative distribution of $\min_j \sup_F - \max_{S' \in F} \min_j$ across decided rows — what fraction of rows the bound understates by at least a given margin. Colour is the declared universe size and linestyle the assumption level (solid A1, dashed A2), with the decided row counts given in the panel. The mass at zero is the majority: on most rows the two quantities agree exactly, and the curve falls off within a few hundredths. Right: the incidence of the sign flip, over the rows the cheap test admits and the solver decided, which is the only population on which the test can err in this direction and the only place in the figure where the two levels are pooled — with thirteen gaps in the whole grid the split would be unreadable. Each bar is labelled (rows where the test asserts a false leader)/(rows the cheap test admitted and the solve decided). The bars are counts from one nested design whose rows are dependent and were selected by solver tractability, so no interval is drawn and no trend is read from the panel: this shows where the errors are, not how their rate depends on universe size.

price of the conjunction: the screen settles each rival separately, and the solve must satisfy them all at once.

J ACQUISITION: A DEPLOYABLE MEASUREMENT RULE

§4.1–§4.3 characterise what a partially observed matrix leaves undetermined. That diagnosis makes a decision problem askable and, on the block where ground truth exists, answerable: *one more evaluation costs money — which one should the platform buy?* This appendix derives a rule that answers it from the evidence the platform already holds, states what it guarantees, proves mechanically that it never reads the outcome it is about to buy, and measures it against nine alternatives, one of which is an outcome-aware one-step greedy comparator rather than a bound.

The decision problem. Fix a target set of comparison claims — here the $\binom{10}{2} = 45$ orderings inside the reported top ten — and write $A(F)$ for the number of them that the feasible set F (§3.1) still leaves undetermined. For a candidate cell $q = (k, b)$ and a candidate value y in its admissible range A_b, let $F(q = y)$ be F with that cell fixed at y. Three one-step objectives are available, and they differ in what they assume:

$$G(q) = A(F) - \max_{y \in A_b} A\big(F(q{=}y)\big) \qquad \text{settled } \textit{whatever} \text{ the measurement returns,} \tag{9}$$

$$C(q) = \min_{y \in A_b} \Big[W(F) - W\big(F(q{=}y)\big) \Big] \qquad \text{guaranteed contraction of total width,} \tag{10}$$

$$\bar{G}(q) = A(F) - \mathbb{E}_{y \sim \mathcal{U}(A_b)} A\big(F(q{=}y)\big) \qquad \text{settled } \textit{on average} \text{ under a prior,} \tag{11}$$

where $W(F) = \sum_p \big(\sup_F d_p - \inf_F d_p\big)$ is the total width of the identified intervals.

equation 9 and equation 10 are worst-case statements and carry a theorem; equation 11 is a modelling choice and is labelled as one throughout. Its prior is stated in full there and nowhere else: the

support is A_b, the cell's own admissible range; the law is uniform on it; and it is fixed before the cell is chosen, from the observed matrix, the mask, the declared rule and the ranges — the same information the two worst-case objectives use, and nothing about the value being bought. It is a declaration about where an unmeasured cell could lie, not a posterior over missing values and not an estimate of them, and this paper nowhere calls a rule that uses it Bayesian. The rule this appendix proposes is lexicographic: it measures the cell maximising equation 9 and breaks the tie among the maximisers on a secondary key. The numbers below are reported for the prior-free form of that rule, G then C, and §J measures it against the other three ways of breaking the same tie — including the one, $\bar{G}$, that needs a declared prior and does better on these runs. The certificate is the primary key's and is the same under all four. That primary key is the objective that is *about* identification rather than about the measurement: it counts the published claims whose label stops depending on the answer.

What the rule reads. The observed matrix, the current mask, the declared aggregation rule, and the admissible ranges. It does not read the withheld value of any cell, it does not consult a running outcome, and it calls no solver. §J proves the first of these mechanically rather than asserting it.

WHY ONE STEP IS EXACTLY COMPUTABLE

Three structural facts, all inherited from the column factorisation of §3.1, make $\max_y$, $\min_y$ and $\mathbb{E}_y$ above exactly computable rather than estimated.

Proposition 4 (Column locality)**.** *$d_p = \rho(i) - \rho(j)$ is a mean over columns of a per-column difference, and F is a product over columns. Fixing cell (k, b) therefore changes every claim's identified interval through column b alone, so the acquisition score of a cell depends on that cell's own column and on nothing else.*

The first half is equation 2 restated; the consequence is what makes the scan cheap, because it means the price of a candidate can be computed without touching the other $|B| - 1$ columns.

Proposition 5 (Finite adversary)**.** *Let $E_b = \{\ell_b, u_b\} \cup \{$values observed in column $b\}$ and let Y_b be E_b together with the midpoint of each pair of consecutive elements of E_b. For fixed (k, b) the map $y \mapsto A(F(q{=}y))$ is constant on each open interval between consecutive elements of E_b, so*

$$\max_{y \in A_b} A\big(F(q{=}y)\big) = \max_{y \in Y_b} A\big(F(q{=}y)\big),$$

and the expectation is exact as well:

$$\mathbb{E}_{\mathcal{U}(A_b)}\big[A(F(q{=}y))\big] = \sum_{y \in Y_b} \frac{|y\text{'s cell}|}{u_b - \ell_b}\, A\big(F(q{=}y)\big),$$

the second exactly, because the prior's mass on an interval is that interval's length and the endpoints of E_b carry zero mass.

The proof is the one already used for the identification side. Each free non-pair cell enters a column's contribution through the piecewise-constant function of §3.3, whose only breakpoints are the pair's two pinned values, and each of those is itself one of two endpoints of $[\ell_b, u_b]$; so as y varies, every term that changes does so only when y crosses a value in E_b. The adversary is therefore not a continuum but a finite list, and the worst case over it is one loop.

Proposition 6 (Exactness and cost)**.** *$G(q)$, $C(q)$ and $\bar{G}(q)$ are computed in closed form from the observed matrix, the mask, the declared rule and the ranges: no solver, no sampling, and no outcome model. The cost of scoring one cell is $O(|Y_b| \cdot P)$ for P target claims, and choosing the best of all free cells is $O(NB\,|Y_b|\,P)$; on the block used here $|Y_b| \le 66$, $P = 45$, and a full scan takes about four seconds.*

The closed form for one slice is what makes this cheap. For a cell (k, b) fixed at y and a claim $p = (i, j)$ not containing k,

$$m(p, k, y) \;=\; E(b, p) \;+\; \frac{c(y; v_i, v_j) - c(x^*; v_i, v_j)}{n - 1},$$

where $E(b, p)$ is the column's contribution with every cell free, v_i, v_j are the pinned pair values, and x^* is the value `best_free_value` gives any free non-pair cell — the same value for every such cell, so no per-cell re-solve is needed. Only the at most nine claims that contain k need the direct evaluation. The implementation (`tools/acquisition.py`) checks this against the definition cell by cell on $4{,}000$ randomised columns with no mismatch, and checks the per-row extreme used for Proposition 8 against a grid brute force on $1{,}112$ columns with no containment violation.

Proposition 7 (The certificate is non-negative, and it accumulates)**.** *If the admissible ranges are frozen for the episode, then $F(q{=}y) \subseteq F$ for every $y \in A_b$, so A cannot rise: $G(q) \geq 0$ and $C(q) \geq 0$ for every cell, and after any sequence of acquisitions*

$$A(F_T) \;\leq\; A(F_0) - \sum_{t=0}^{T-1} G(q_t).$$

So the trajectory is certified in advance*: the platform knows, before it buys anything, a lower bound on how many claims the budget will settle.*

The nesting argument is what a guarantee needs, and it is exactly what forces the next subsection's design choice. It also makes the objective honest about its own limits: on the runs below the bound is positive but small — $\sum_t G(q_t) = 42$ against a realised drop of 123 for the proposed rule — because a guarantee can only count what no completion can overturn. §J measures how much of the realised drop that leaves to the tie-break.

Proposition 8 (A leader-set bound, for free)**.** *The same column factorisation bounds the set of models that could be the leader. With u_k and ℓ_k the column-wise extremes of row k's win rate over the admissible completions, every completion that puts k on top satisfies $u_k \geq \max_j \ell_j$. The set $\{k : u_k \geq \max_j \ell_j\}$ therefore* contains *the possible-leader set of §4.3 and is computed with no solver. It is an outer bound, not the exact set, and is reported as such; it contracts from a mean of 13.2 models to 9.7 under the proposed rule and to 8.6 under the expected-gain rule.*

The range contract, and exactly what the certificate assumes

Proposition 7 has a hypothesis, and the hypothesis is about *where the measurements land*: $F(q{=}y) \subseteq F$ needs $y \in A_b$. Nesting is therefore a statement about the ranges, not about the analysis, and it cannot be a guarantee unless the ranges are fixed in a way the data cannot contradict. This subsection states which of the three levels has that property, what the platform holds when it does not, and what it must do when the assumption is falsified. The measured quantities that follow from it are given in §J.

A0 is unconditional, where the platform declares a domain, and it is per column. At A0 the range *is* the metric: a metric the release defines as a fraction of instances lies in $[0, 100]$, and no measurement, present or future, observed or withheld, can fall outside it. So at A0 the hypothesis of Proposition 7 holds by construction,

$$A(F_T) \;\leq\; A(F_0) - \sum_{t=0}^{T-1} G(q_t)$$

unconditionally, with no falsification branch, and $G(q) \geq 0$ for every cell. This is the range to declare when a platform is willing to declare one. It is also the weakest of the three levels: A0 admits every completion A1 and A2 admit, so it identifies the fewest claims on the same evidence. Weakest is not none, and this is the sense in which A0 is not assumption-free: it *is* an assumption, that the release's declared domain for a metric is the set that metric can take. What distinguishes it from A1 and A2 is not that it assumes less but that this assumption is about the metric rather than about the sample, so the data cannot contradict it — and where it can, as on the three columns below, it is not asserted at all.

The declaration is a property of the *metric*, so it is a property of each column separately, and it is available for a column only if the release's own definition of that metric bounds it. In this block 19 of the 22 columns are bounded and 3 are not: the bits-per-byte columns (The Pile, TwitterAAE, ICE) are defined by the release as the average number of bits per byte according to model probabilities — a mean of $-\log_2 p$ over bytes, non-negative by definition, with no maximum the release declares.

Oriented, such a column lies in $(-\infty, 0]$, so its A0 range is a half-line: $[0, 100]$ is not its declared domain, and negating the block's percent record to $[-100, 0]$ is not either — that asserts a loss of at most 100 bits per byte, which no source states and a withheld value could falsify. A0's claim to be unfalsifiable would be false on those three columns, so rather than assert it the columns are *outside the A0 universe*: the A0 study of §4.4 runs on the 19 columns the release bounds, which is a restriction of the study and is reported as one. The two ground-level reasons a stand-in endpoint would not have worked are in Appendix J; briefly, the identification code probes each region's interior by midpoints ($\frac{1}{2}(lo + hi)$), and with $lo = -\infty$ every such midpoint *is* $-\infty$, so the region below a column's first observation is never evaluated at all, and the expected-gain prior of §4.4 has no total mass on a half-line.

A1 and A2 are a frozen contract on the observed data, and they are falsifiable. A1's range is $[\min_b, \max_b]$ of the *observed* cells of the column. A2 cuts that range to the half-span about the observed median, with med_b the median of column b's observed cells and $h_b = (\max_b - \min_b)/4$:

$$A2_b \;=\; \big[\max(\min_b,\ \mathrm{med}_b - h_b),\ \ \min(\max_b,\ \mathrm{med}_b + h_b)\big], \tag{12}$$

which is the whole of A1 exactly when the observed median lies within a quarter-span of the midpoint, and a strict sub-interval otherwise. It is not a platform declaration and it is not a claim about the metric; it is a statistic of the data, frozen at the start of the episode and held fixed thereafter. Nothing prevents a withheld score from lying outside it, and when one does, the slice $F(q{=}y)$ is *empty*: the assumption, not the analysis, has been contradicted. Three things are then true, and are reported separately.

(i) The certificate is segmented. Split the episode at its falsification steps. For a maximal run of consecutive steps $a..b$ that did not widen, every slice within the run is nested, so

$$A(F_b) \;\le\; A(F_{a-1}) - \sum_{t=a}^{b} G(q_t). \tag{13}$$

Writing the non-widening steps as N and the widening steps as W gives the global statement

$$A(F_T) \;\le\; A(F_0) - \sum_{t\in N} G(q_t) \;+\; \sum_{t\in W} \delta_t, \qquad \delta_t \;=\; A(\tilde{F}_{t-1}) - A(F_{t-1}), \tag{14}$$

where $\tilde{F}_{t-1}$ is the pre-step state with the range already stretched to admit the value about to be revealed and the cell still free. The correction is therefore the widening's *own* damage, on the old partition, and not the realised rise across the step: the realised change is (claims the reveal settles) minus δ_t, so it is negative whenever the reveal settled more than the widening cost, and a correction built from it understates the damage at exactly those steps. The gain sum runs over N only, which is the content of the whole formula: a widening step contributes no guarantee and must not be counted as if it did. Summing over all steps instead and adding back what was claimed at the widening ones gives the equivalent form $A(F_0) - \sum_t G(q_t) + \sum_{t\in W}(G(q_t) + \delta_t)$. equation 13 and equation 14 are what this paper verifies per step; the unsegmented inequality is not verified at A1, because its hypothesis is not true there.

Both forms are proved as algebra over arbitrary integers, with the two near-miss variants refuted by minimal counterexample, in `paper/_certcheck.py` (listed in `paper/checks.md` with the rest of the artifact checkers): the theorem is G summed over N with the δ correction, and the two forms that have been used elsewhere — G summed over *all* steps with only the δ correction, and the realised-rise correction $\sum_{t\in W} \max(0, A(F_t) - A(F_{t-1}))$ — both fail, the first by exactly the guarantee claimed at the widening steps and the second by the settled claims at those same steps. The algebra cannot supply the bridge from the formula to the run; that bridge is the measured per-segment inequality below.

(ii) Falsification is detected at purchase time, and the response is defined. The platform sees the revealed value as it buys the cell, so it always knows whether it is still inside a segment. On a reveal outside A_b it widens A_b to contain the revealed value, rebuilds F, and restarts the chain from that step. It held a valid certificate up to the falsification and holds a valid one from that step on; what it cannot do — and what an unqualified "certified" would wrongly claim — is carry the pre-falsification certificate across it. Widening rather than discarding is forced: leaving the cell outside the range makes F empty and every label vacuous, as Table 9 shows.

(iii) The falsification rate is a property of the level, and is measured. Because A0 cannot be falsified and A1 can, an episode's certificate at A1 is exactly as long as its range-valid prefix. §J reports that prefix, the number of falsifying reveals, and how many of the twelve episodes stay range-valid throughout.

Table 9 gives a measured instance of the mechanism, found by search over small matrices rather than constructed by hand.

WHAT THE RANGE CONTRACT COSTS, MEASURED

The four quantities the contract owes a reader are measured, not asserted: how many acquired outcomes violate the initial range, how many episodes stay valid throughout, how many steps a valid certificate covers, and what the widening does when it happens.

The declared domain is never violated, and it is declared per column. Every acquired value in every arm and every run was checked against the oriented declared domain of *its own* column. **Zero** of the $2{,}400$ acquisitions fell outside it, on arms whose schedules differ from one another and therefore probe $2{,}400$ different cells. The unconditional inequality of Proposition 7 is not being asserted about a range the data might contradict; on this grid the data did not contradict it. The second pass at A0 (`results_a0.json`) runs the same twelve instances on the 19 columns the release declares a bounded domain for — re-deriving the reported head inside that universe, so its claim counts are that sub-study's own and are not comparable with the A1 column of Table 11 — and reports **zero** of its $2{,}400$ acquisitions outside the declared domain, 72 of 72 deployable arm-runs range-valid throughout, and a correction term of 0: the one row here where the widening term is zero by construction rather than by measurement. Its counts are given in Appendix J.

The frozen A1 range is violated often, and that is the point. 158 of the $2{,}400$ acquisitions (6.6%) returned a value outside the frozen A1 range of its column: 101 for the six deployable rules out of $1{,}440$ acquisitions, 34 for the three heuristics out of 720, and 23 for the outcome-aware arm out of 240. A range that is a statistic of the observed cells is falsified by a withheld cell sooner or later, which is why the guarantee is not stated on it unconditionally.

Whole episodes that stay valid are the exception. Counting an episode as range-valid when it never widens: `gain-contract` 1 of 12, `contract-gain` 2 of 12, `gain-rand` 4 of 12, `gain-widest` 3 of 12, `expgain` and `gain-expgain` 1 of 12 each — 12 of the 72 deployable arm-runs. On the declared domain the same count comes out 72 of 72: the A0 pass is measured, and it could not have come out otherwise, since no reveal there can leave the range.

The segmented certificate covers most of the budget anyway. Splitting each episode at its widening steps, the deployable rules' segments cover $1{,}339$ of $1{,}440$ steps (93.0%), and **all** 102 **scored segments satisfy** equation 13 — the gain the rule read off the state before choosing never exceeded the drop the run actually realised inside that segment. The 101 steps that are not inside a segment are exactly the falsifying purchases, and their effect is the correction term of equation 14. That term is not zero on this grid, and it is reported in the form the theorem states: for the proposed rule the guarantee claimed at a step the reveal then falsified — where the rule's maximisation ranged over a partition the value then left — totals 8, and the undetermined claims the widening itself re-opened total 0. The certified reduction in undetermined pairs is therefore 34, not the raw $\sum_t G(q_t)$ of 42: the 8 points claimed at the steps that were then falsified are exactly what equation 14 refuses to count. Over the four guaranteed arms and all 48 arm-runs the two terms total 24 and 1. The realised-rise variant of the same correction reads 0 in every one of those runs, which is the measurement of why it is not the theorem: the reveal at a falsifying step also settles claims of its own, so the net step-to-step change can stay at zero while the cost that step imposed on the bound is still there. The initial unbroken chain — what a platform holds if it never re-plans — is shorter: for the proposed rule it has mean length 5.4 of 20, and in three runs it is empty.

The label counts never rise, which is what the correction term rules out. The undetermined count rose at no step of any arm in any run, and it fell strictly in 116 of the 120 arm-runs — the four exceptions are one budget mask on which `gain-rand` and `random` settled nothing and so moved nothing. This is the measured content of the design choice at the top of this subsection: the episode widens the range and fixes only the acquired cells, so the feasible set grows no smaller between

segments and no label regresses. The dynamic alternative in Table 9 is what happens when it is refit instead.

The state a platform is in after a falsification is therefore: the pre-falsification certificate is still valid but stops where it stopped; F has been rebuilt with the revealed value inside its range; the cells already bought are fixed; and the next purchase is scored on the new state. What it may not do is quote the pre-falsification inequality as if the chain had never broken.

What the same study reports on the declared domain. The A0 pass is not a second view of the numbers above. It is a study of the 19 columns the release declares a bounded domain for, and the reported head is re-derived inside that universe, so the two passes are not comparing the same 45 comparisons: the A0 head differs from the A1 run's in nine of the twelve instances, by one model in eight of them and by two in one, and the two head sets share between 24 and 42 of the 45 orderings they are asked about. On that universe the proposed rule establishes 21.2 of 45 comparisons on average, pure expected gain 23.7, the outcome-aware comparator 23.4, the widest-pair heuristic 19.3 and random 12.9; the correction term is 0 and the certificate covers all 20 steps of all 72 deployable arm-runs, which is what an unfalsifiable range means. The orderings agree with the A1 ones; the two levels' counts are different objects — a different column universe and, because the head is recomputed on it, a different target set — and this paper reports them side by side rather than differencing them.

THE RULES COMPARED

Ten rules allocate 20 acquisitions each, 2.8% of the 704-cell block.

Every rule is named by the key it maximises, in the order it uses them, because four of the ten share a primary key and the difference between them *is* the tie-break. We write $G(q) = A(F) - \max_y A(F(q{=}y))$ for the *certified* gain — the claims settled whatever the measurement returns — $C(q) = \min_y [W(F) - W(F(q{=}y))]$ for the certified contraction of total interval width, and $\bar{G}(q) = A(F) - \mathbb{E}_y A(F(q{=}y))$ for the *expected* gain. G and C are worst-case over the outcomes the range admits, so a rule that reads only them uses no distribution over outcomes at all.

(G, C) (the arm key `gain-contract`) is the rule this paper proposes, and it is prior-free: both keys are certified quantities read off the feasible set, and the choice is blind to the value about to be revealed. (C, G) (`contract-gain`) is the other certified rule, maximising the guaranteed contraction; it reads the same information and differs only in the objective, and is included because a myopic count can be blind to progress one measurement away from settling a claim. $(\bar{G}, G)$ (`expgain`) is pure expected gain, $\arg\max_q \bar{G}(q)$, and it is not prior-free. It is also deliberately *not* called Bayesian: the prior is a declared *uniform* distribution on the cell's own admissible range $[\ell_b, u_b]$, fixed before the cell is chosen, so the rule remains blind to the withheld value. The prior is the modelling assumption and nothing else is — the support is the frozen range, not the release's declared domain and not the observed values, and the density is $1/(u_b - \ell_b)$, so $\bar{G}(q)$ is the mean of $A(F(q{=}y))$ over y drawn uniformly from that interval. It uses exactly the information the certified rules use. It is reported as a modelling choice and not as a theorem, and it is not defined on an unbounded range, which is one of the reasons the A0 universe of §J is restricted to the bounded columns. $(G, \bar{G})$ (`gain-expgain`) levels the two families: the certified gain first, the expected gain as tie-break. It is identical to (G, C) wherever some candidate has $G(q) > 0$, so the two differ only in the regime the guarantee cannot rank, which is where the tie-break decides.

Two further arms exist to separate the guarantee from the tie-break, because on most steps the guarantee is zero and cannot rank anything (§J). (G, rand) and (G, widest) share the primary key $\arg\max_q G(q)$ with (G, C) and $(G, \bar{G})$ and differ only in what breaks the tie: the first picks uniformly among the maximisers, the second takes the maximiser whose model carries the widest undecided target interval. The tie-break already inside (G, C) is $C(q)$, the guaranteed contraction, and the one inside $(G, \bar{G})$ is $\bar{G}(q)$; with those two that makes four rules with the same primary key, so comparing them isolates the secondary key with the primary held fixed.

widest takes a missing cell of the target pair whose identified interval is currently widest; **complete** takes a cell of the column with the fewest remaining gaps, ranked by the contribution width that column carries per remaining gap; **random** draws uniformly from the missing cells. These three are

kept from an earlier version of this study so that the new rules are compared against the heuristics they replace; all three read only the published matrix.

oracle (the outcome-aware one-step greedy comparator) scores every candidate by the resolution it *actually* delivers, which requires reading the value about to be revealed. It is optimal at the single step it is standing on and has no claim beyond that. It is *not* an upper bound on a twenty-step episode, and this appendix does not use it as one — a one-step-optimal move can be worse in sequence, the runs below contain arms that beat it, and the only general episode-level upper bound on the objective is the trivial $A(F_0)$, the number of claims that were undetermined to begin with, which every policy attains only by observing everything.

Every rule runs on the same frozen-range state, so the comparison is between objectives and not between assumptions.

No deployable rule reads a withheld value. A checker (`paper/_leakcheck.py`) builds the initial state twice from two matrices that agree on every observed cell and differ at every withheld one, and then asks whether anything a deployable rule can see has moved. Three probes are used — uniform in the range, deliberately out of range, and constant — over the exhaustive set of free cells of one instance and over three instances for the choices.

RESULTS

Every comparison established by every arm, outcome-aware arm included, is *true* against the withheld measurements — 0 unsound of 120 arm-runs — which is the property a platform needs: the labels it would add stay sound as the evidence grows.

The rules that use a secondary key beat every heuristic; the primary key alone does not. `gain-contract` establishes 19.83 of 45 orderings against 17.83 (`widest`), 15.50 (`complete`) and 13.00 (`random`), and the expected-gain rule reaches 21.58. The same holds for `contract-gain` (19.58) and `gain-expgain` (21.33). The exceptions are informative. `gain-widest` reaches 17.67 — above `complete` and `random`, below `widest` — and `gain-rand`, which has `gain-contract`'s primary key and a uniform tie-break, reaches only 14.83 and beats `random` alone. §J shows that this is not an anomaly but the shape of the whole result.

The outcome-aware arm is a comparator, not a ceiling. `expgain` and `gain-expgain` each beat the one-step greedy comparator on 6 of the 12 runs, and `expgain`'s worst run (14) is better than the comparator's worst run (13). A policy that reads the value it is about to buy and picks the best single move is not an upper bound on a twenty-step episode, and the table is the demonstration rather than an assertion: the "percent of the distance to the oracle" figure quoted in an earlier version of this study has been removed, because no such distance is defined.

Neither objective dominates, and the choice between them is not made here. The expected-gain rule is ahead on the mean (21.58 against 19.83) and on the worst run (14 against 10), and it also shrinks the possible-leader bound further (8.6 against 9.7). The guaranteed rule is the one this appendix proposes, because it is the one Proposition 7 certifies and because its primary key is the certified quantity; the expected-gain rule is reported beside it rather than chosen over it, since nine designed instances do not rank two rules that trade one criterion against another.

WHAT THE GUARANTEE CONTRIBUTES, AND WHAT THE TIE-BREAK DOES

The certificate is small, so the honest question is what the ranking above is actually made of. Four rules share the primary key $\max_q G(q)$ and differ only in what breaks a tie, which isolates the secondary key with the primary held fixed. Reading their 240 acquisition steps:

Four things follow, and the last one is the positioning of the whole contribution.

On most steps the guarantee cannot rank anything. For the proposed rule $\max_q G(q) = 0$ over every candidate on 198 of 240 steps (82.5%), and the tie-break therefore fixes the cell on 222 of them (92.5%) — classes 3 and 4 together — leaving 18 steps (7.5%) where the guarantee alone determines the choice. The other three rows are the same: 228, 223 and 229 tie-broken steps.

The realised drop is not where the certificate is. $\sum_t G(q_t) = 42$ against a realised drop of 123: the certificate accounts for 34% of what the run actually settled. Of the rest, 41 comes from steps where $G > 0$ but tied and the tie-break decided, and 54 — the largest single component — from the 198 steps where the guarantee was identically zero and could not have contributed at all.

Holding the primary key fixed, the tie-break moves the result by more than the primary key does. The ablation ladder is `random` $13.00 \rightarrow$ `gain-rand` 14.83 ($+1.83$ from adding the certified primary key to a uniform tie-break) $\rightarrow$ `gain-contract` 19.83 ($+5.00$ more from changing only the tie-break). Over the four rows the certified total $\sum_t G$ is *not* monotone in the realised drop — $29, 37, 42, 21$ against $63, 99, 123, 142$ — so the magnitude of the guarantee does not order these arms. The rule that certifies least settles most.

So the guarantee is a certification layer, not the source of the ranking. Its empirical content is the certificate of §J: a lower bound the platform holds before it buys, valid on the declared domain unconditionally and per segment on a frozen A1 range, verified on 102 of 102 segments, with the widening correction of equation 14 measured at 25 over the guaranteed arm-runs rather than assumed to be 0. Its contribution to the ordering above is the $+1.83$ the primary key buys over a uniform tie-break, and the rest is the secondary key. This is what the earlier version of this study got wrong by attribution rather than by arithmetic: it reported the certified rule's advantage over the heuristics as if the certificate produced it. The correct statement is narrower and still useful, and it is lexicographic. *First maximise the guaranteed gain G; among the candidates that attain it, maximise the secondary key.* The first key is the only part with a proof — it carries the certificate — and on this evidence it is the smaller part of the effect: $+1.83$ of the $+6.83$ over random. The second key carries the rest, and the runs select it. Of the four secondary keys measured, $\bar{G}(q)$ reaches 21.33 with a worst run of 14; $C(q)$ reaches 19.83 with a worst run of 10; the widest-interval key 17.67 and the uniform key 14.83. $\bar{G}$ rests on a declared prior and C does not, so the prior-free form of the rule is G then C, and a platform willing to declare the prior should break ties on $\bar{G}$ and take the $+1.50$ mean the runs show. The certificate is the same under either: it is the primary key's, not the tie-break's, and it does not depend on how the tie was broken. That gives the division of labour a measurable form, and the measurement supports it: ranking purely on $\bar{G}$ with no primary key reaches 21.58 with a worst run of 14, and the lexicographic rule that carries the certificate reaches 21.33 with the same worst run, so the guarantee costs 0.25 of the mean and nothing at the worst run. The certified layer supplies the certificate; the expected-gain layer supplies the efficiency subject to it.

What this does not establish. The rule is myopic — one step, not a plan — and the runs do not show that any rule is optimal, that the ordering of the heuristics transfers to other instances, or that the expected-gain rule's mean advantage would survive a wider instance set. The twelve runs are nine designed instances the matrix happened to supply, not a sample, so every number above describes them. One designed discriminator failed and is reported rather than dropped: each instance was seeded with a target pair undetermined in every one of the twelve, and twenty acquisitions resolve it in exactly one arm-run out of the hundred and twenty, on that run's nineteenth acquisition. A pair the budget never reaches says nothing about the rules, so it is not used to rank them; the seeding is kept because it removes the pair from the target count identically for every arm. What is established is narrower than the earlier version of this study claimed and better founded: a rule a platform can run on the evidence it already publishes, provably blind to the outcomes it is buying, whose certificate is valid under a stated range contract and whose limits under that contract are measured rather than assumed, and which beats the heuristics on the block where the truth is known — with the part of that advantage that comes from the certificate, and the part that does not, reported separately.

K Extended related-work comparisons

Partially identified rankings from pairwise interactions. The nearest neighbour on the phrase "partially identified rankings" is Crippa & Fedchenko (2024), who characterise the identified set of a ranking from incomplete *pairwise interaction* data under a link function. Three differences matter. Their unknown is a latent merit and their data are noisy draws from a link function, so their identified set is a statistical object with moment inequalities and asymptotic tests; ours are deterministic unobserved measurements with no sampling model. Their incompleteness is a missing edge of a tournament graph; ours is a missing cell of a score matrix, constrained by the benchmark's

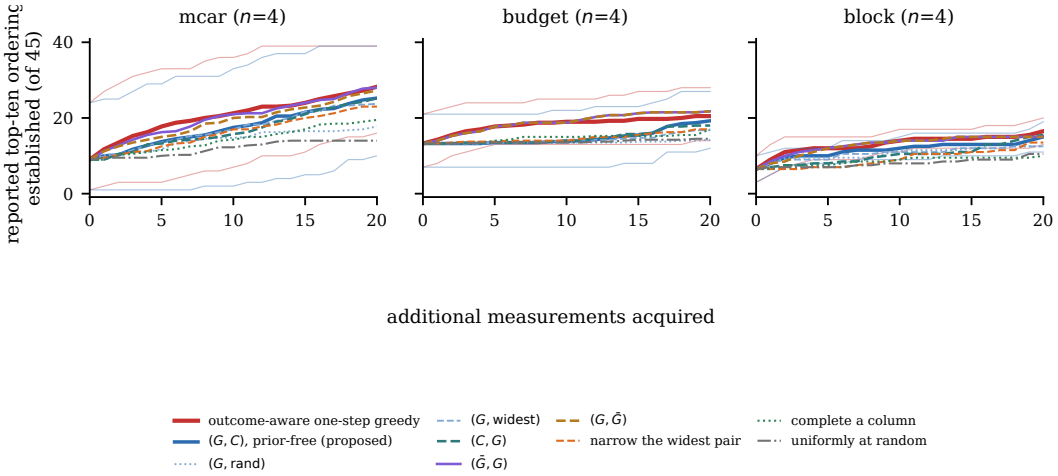

Figure 6: Comparisons established inside the reported top ten as measurements are added, by hiding pattern. Lines are means over the instances of that pattern and the thin rails above and below them span the instances' range (minimum to maximum). The `oracle` curve is the outcome-aware one-step greedy comparator: it reads the outcome and is optimal at the single step it stands on, and it is a comparator, not a bound on the episode. Every other curve is deployable and never reads it. The legend names each rule by the keys it maximises in the order it uses them: (G, C) is the proposed prior-free rule, $(G, \bar{G})$, $(\bar{G}, G)$ and (C, G) are the other three objectives, and (G, rand) and (G, widest) share (G, C)'s primary key and differ from it only in the tie-break, so the four G-keyed curves are drawn as variants of the proposed rule's colour; the remaining three are the heuristics the earlier version of the study compared. The target set is fixed at the reported head of the *initial* masked instance, so the curves measure how much of what the platform published the evidence comes to settle.

admissible range. And their object is the ranking itself, obtained pairwise, whereas the leaderboard claim is set-valued over the whole matrix at once — the distinction that §3.3 shows is not cosmetic.

Elicitation and active evaluation, in detail. Choosing the comparison to buy next is the query-selection problem of preference elicitation, and that problem is old and well developed: Pini et al. (2007) give complexity results for determining the winners of several aggregation rules from incomplete preference data, Lu & Boutilier (2020) choose queries to call a winner while hedging against the preferences they have not seen, and Shah & Yu (2026) do it with next-best queries under a competitive criterion. The same move appears in evaluation as adaptive benchmarking: Mohankumar & Khapra (2022) choose which system pairs to compare, Li et al. (2025) which benchmark items to run, and Arviv et al. (2026) when to stop.

Tables 13 and 14 record the comparison on eight dimensions for each of the six, and each cell was read from a source rather than inferred from a citation. Five were read as full text: the four evaluation and elicitation preprints (Shah & Yu, Mohankumar & Khapra, Li et al., Arviv et al.), and Pini et al., whose cells here rest on the openly available *conference* paper (Pini et al., 2007). The journal version (Pini et al., 2011) is the version of record for the same work but is paywalled (its DOI returns 403) and was not inspected; the two rows citing Pini et al. therefore cite the conference paper for every cell, and no cell is attributed to the journal article. The sixth, Lu & Boutilier, is not openly available in any version: only the publisher's abstract was inspected, which is what the dagger on its rows records. That abstract states the partial-ranking input, the minimax-regret objective, the preference queries to voters and the tractability split, and those are the only cells populated from it. Cells the abstract does not establish are marked "not established from the inspected source" rather than filled in from general knowledge about regret-based elicitation — a minimax-regret method is worst-case over the completions consistent with the partial rankings by definition, but the abstract does not say so, and the table does not either. Two cells were corrected when the sources were read against this table: the Li et al. row had credited their acquisition policy's baselines to the policy, and the Arviv et al. row had narrowed their scope to single-model scoring. The readings, with the passage each cell

rests on, are recorded in the research ledger; the preprints themselves are not redistributed with the repository. Four differences matter for what §4.4 claims.

The unknown. Theirs is a preference relation, or an unobserved item outcome, that the query *reveals*; ours is a benchmark score constrained to a declared interval. A query therefore buys a narrowing of an identified set rather than the observation of a preference, and what it settles is a set-valued property of the completion (§3.3), not a realised sample.

The objective. Theirs is a decision — the winner, the ranking, the stopping time; ours is a count of *already published* claims that the evidence comes to settle. The distinction is not verbal: a rule optimising for the decision may leave published claims undetermined, and a rule optimising for the claims may never call the winner, which is exactly the trade the runs in Appendix J measure. It is also the distinction the last column of Table 14 records: none of the six takes an already-published comparison as the thing to be resolved.

The guarantee. Their analysis bounds query complexity, or expected loss under a model of the answer. Ours is a worst case over completions with no answer model, and the acquisition's own guarantee (Proposition 7) is a lower bound on settled claims that the platform holds *before* running the query — and, because the bound is read off the state, it is conditional on the range contract of §J in a way that none of the six's guarantees is.

Exactness. Because the aggregate is a column-wise mean and the feasible set is a box, the worst case over completions is a finite list (Proposition 5) and the score is in closed form. The general elicitation problem offers no such reduction — determining what partial preference data settles is itself a complexity question with hardness results for several rules (Pini et al., 2007) — and it is that reduction, not the idea of scoring a candidate query, that the certificate rests on. Query selection by a score read off the current state is not new here, and neither is adaptive evaluation; what is new is the pairing of the published-claim objective with this exactness, which is what turns a guarantee from an expectation into a certificate.

What the complete-case restriction buys and costs. Gordienko et al. (2026a) demonstrates directly what that restriction buys and costs: adding a single model to a HELM MMLU comparison changes the top of the aggregate ranking in 44 of 57 subjects.

Estimand-aware comparison, in detail. Huang (2026) shows that what a leaderboard comparison means depends on the estimand — task mixture, label source, pass@k, cost — and proposes a common-support check within a benchmark. We share the motivation that a reported ordering may not compare what it appears to compare, and extend the object: from within-benchmark estimand choices, where the data are complete, to across-benchmark coverage, where they are not. Their common-support check is in fact the coarsest resolution of our labelling — support present versus absent, with no use of the admissible ranges that let a claim be established despite missing cells.

Entry-specific bounds and the case for a smaller submatrix. Entry-specific bounds under non-uniform sampling show that it is "often better and sometimes optimal to run estimation algorithms on a smaller submatrix rather than the entire matrix" (Xi et al., 2024). That is a genuine argument for the deletion practice and it deserves a direct answer. The result is about *estimation risk under a sampling model*. Our question carries no sampling model: we quantify over every completion in the admissible range, not over a distribution of completions, so removing observations cannot add information — it can only shrink the claim being made. §4.1 shows precisely that: deletion turns an honest "undetermined" into a confident "established" about a smaller question.

Benchmark operators as social choice. The closest work in *platform* treats the aggregation rule itself as the object: for HELM's MMLU slice, Gordienko et al. (2026b) recast benchmark-specific training as shift bribery under arithmetic mean, median, *mean win rate* and pairwise majority, and a companion paper asks when an aggregation operator is coherent and stable at all (Gordienko et al., 2026a). Their question is which rule resists manipulation; ours is what an incomplete matrix establishes under a given rule. Both work from a *complete* matrix obtained by keeping only the models with scores on every retained task — the common-support remedy whose cost §4.1 measures. We claim no novelty for the rule's dependence on the model set, only for its consequence: because the

comparison set is whatever was observed, a missing cell changes its column's normalisation, so "i leads" is a statement about a family of matrices whose normalisations move with the completions.

Other partial-data ranking settings. Non-identifiability of unobserved consideration sets in Plackett–Luce ranking (Aoki-Sherwood et al., 2024) is a further instance of rankings that are only partially determined, with the unobserved object being a choice set rather than a cell of a score matrix.

L THE ESTIMAND, IN ONE PANEL

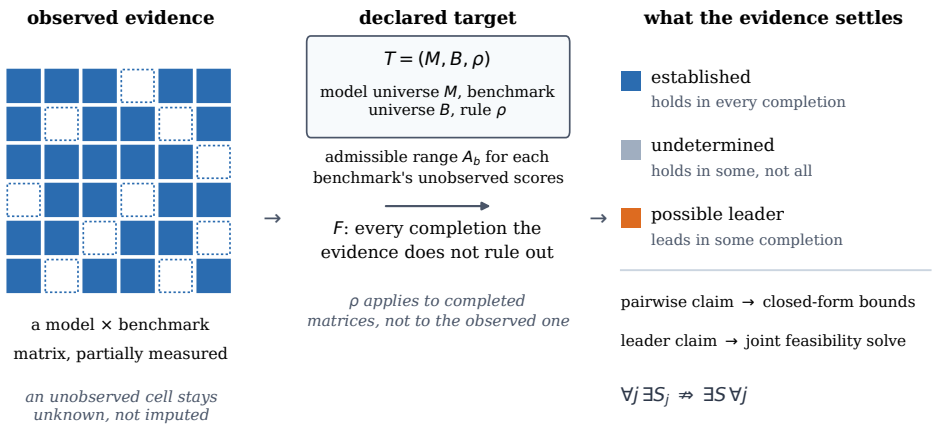

Figure 7: The paper in one panel. A platform supplies a declared model universe, a declared benchmark universe and an aggregation rule — three things we take as given — and applies its rule to the partial evidence it holds. The estimand this paper studies is that *same* rule read on the matrix completed over the declared universe: a counterfactual reading, and the only ingredient we choose. Each unobserved cell is held only to an admissible range, and a claim is labelled by whether it holds in every completion those ranges leave open, in some, or in none. Pairwise claims and established leaderhood follow from closed-form bounds; possible leaderhood needs a joint solve.

The figure carries no experimental number: every quantity in it is a definition, which is what makes it the place a reader can check the paper's logic rather than its results. The steps it draws are the ones §3.1–§3.3 formalise.

higher in the reported order	lower	inf A0	inf A1	inf A2	labels
Llama 2 (70B)	LLaMA (65B)	-0.35417	-0.34848	-0.30824	UUU
Llama 2 (70B)	text-davinci-002	-0.33144	-0.32860	-0.30256	UUU
Llama 2 (70B)	Mistral v0.1 (7B)	-0.33759	-0.33191	-0.29167	UUU
Llama 2 (70B)	Cohere Command beta (52.4B)	-0.29593	-0.29309	-0.26705	UUU
Llama 2 (70B)	text-davinci-003	-0.30682	-0.30398	-0.27699	UUU
Llama 2 (70B)	Jurassic-2 Jumbo (178B)	-0.26847	-0.26562	-0.23958	UUU
Llama 2 (70B)	Llama 2 (13B)	-0.30066	-0.29498	-0.25473	UUU
Llama 2 (70B)	TNLG v2 (530B)	-0.22254	-0.21970	-0.19366	UUU
Llama 2 (70B)	gpt-3.5-turbo-0613	-0.27746	-0.27178	-0.23059	UUU
LLaMA (65B)	text-davinci-002	-0.35322	-0.35038	-0.32434	UUU
LLaMA (65B)	Mistral v0.1 (7B)	-0.35938	-0.35369	-0.31345	UUU
LLaMA (65B)	Cohere Command beta (52.4B)	-0.31771	-0.31487	-0.28883	UUU
LLaMA (65B)	text-davinci-003	-0.32765	-0.32481	-0.29877	UUU
LLaMA (65B)	Jurassic-2 Jumbo (178B)	-0.29025	-0.28741	-0.26136	UUU
LLaMA (65B)	Llama 2 (13B)	-0.32244	-0.31676	-0.27652	UUU
LLaMA (65B)	TNLG v2 (530B)	-0.24337	-0.24053	-0.21449	UUU
LLaMA (65B)	gpt-3.5-turbo-0613	-0.29830	-0.29261	-0.25237	UUU
text-davinci-002	Mistral v0.1 (7B)	-0.18892	-0.18608	-0.10275	UUU
text-davinci-002	Cohere Command beta (52.4B)	-0.08286	-0.08286	-0.01373	UUU
text-davinci-002	text-davinci-003	-0.06629	-0.06629	-0.06629	UUU
text-davinci-002	Jurassic-2 Jumbo (178B)	+0.03741	+0.03741	+0.03741	EEE
text-davinci-002	Llama 2 (13B)	-0.15388	-0.15104	-0.06771	UUU
text-davinci-002	TNLG v2 (530B)	+0.06061	+0.06061	+0.06061	EEE
text-davinci-002	gpt-3.5-turbo-0613	-0.12879	-0.12595	-0.04261	UUU
Mistral v0.1 (7B)	Cohere Command beta (52.4B)	-0.33617	-0.33333	-0.30729	UUU
Mistral v0.1 (7B)	text-davinci-003	-0.34612	-0.34328	-0.31723	UUU
Mistral v0.1 (7B)	Jurassic-2 Jumbo (178B)	-0.30871	-0.30587	-0.27983	UUU
Mistral v0.1 (7B)	Llama 2 (13B)	-0.34091	-0.33523	-0.29498	UUU
Mistral v0.1 (7B)	TNLG v2 (530B)	-0.26184	-0.25900	-0.23295	UUU
Mistral v0.1 (7B)	gpt-3.5-turbo-0613	-0.31487	-0.30919	-0.26894	UUU
Cohere Command beta (52.4B)	text-davinci-003	-0.07813	-0.07813	-0.07813	UUU
Cohere Command beta (52.4B)	Jurassic-2 Jumbo (178B)	+0.00189	+0.00189	+0.00189	EEE
Cohere Command beta (52.4B)	Llama 2 (13B)	-0.18845	-0.09422	-0.07907	UUU
Cohere Command beta (52.4B)	TNLG v2 (530B)	+0.04877	+0.04877	+0.04877	EEE
Cohere Command beta (52.4B)	gpt-3.5-turbo-0613	-0.16430	-0.07008	-0.05492	UUU
text-davinci-003	Jurassic-2 Jumbo (178B)	-0.05729	-0.05729	+0.01184	UUE
text-davinci-003	Llama 2 (13B)	-0.17945	-0.11127	-0.02746	UUU
text-davinci-003	TNLG v2 (530B)	+0.00000	-0.00000	+0.03504	UUE
text-davinci-003	gpt-3.5-turbo-0613	-0.15341	-0.08523	-0.00142	UUU
Jurassic-2 Jumbo (178B)	Llama 2 (13B)	-0.21591	-0.18987	-0.10606	UUU
Jurassic-2 Jumbo (178B)	TNLG v2 (530B)	-0.04403	-0.04403	-0.04403	UUU
Jurassic-2 Jumbo (178B)	gpt-3.5-turbo-0613	-0.19176	-0.16572	-0.08191	UUU
Llama 2 (13B)	TNLG v2 (530B)	-0.29782	-0.29498	-0.26894	UUU
Llama 2 (13B)	gpt-3.5-turbo-0613	-0.35180	-0.34612	-0.30587	UUU
TNLG v2 (530B)	gpt-3.5-turbo-0613	-0.23769	-0.23485	-0.12784	UUU

Table 2: Every pairwise ordering inside the reported top ten of HELM's declared `classic: core_scenarios` Accuracy universe, one row per comparison, on the recovered full-precision matrix. The first two columns are the two models in the order the platform reports them; the next three are the infimum over admissible completions of the aggregate difference (higher minus lower) at A0, A1 and A2, in win-rate units; the last column gives the label at A0, A1 and A2 (**E** established, **U** undetermined, **R** refuted). The label is **E** exactly where that infimum is strictly positive — an infimum of zero is a tie, which establishes no strict order — so the entries printing -0.00000 and $+0.00000$ are **U**, not **E**. Counting **E** down the last column reproduces Table 1: 4 at A0, 4 at A1, 6 at A2, and no **R** anywhere. The four comparisons established at A0 and A1 are the four whose lower model is `Jurassic-2 Jumbo (178B)` or `TNLG v2 (530B)`; the two that become established only at A2 are `text-davinci-003` over the same two models, whose A1 infimum is -0.00000 — a tie read as a non-establishment, not a reversal.

restriction	block	cells kept	discarded	est. /45	leader	margin vs. runner-up
as reported	67×16	876	—	4	`Llama 2 (70B)`	-0.348
common support (columns)	67×7	469 (53.5%)	9 scenarios	45	`Llama 2 (70B)`	$+0.0152$
no missing task (rows)	29×16	464 (53.0%)	38 models	45	`text-davinci-002`	$+0.0100$

Table 3: The two deletion remedies on the declared universe of §4.1, and the declared universe itself. "est." counts the reported top-ten orderings that come out established, at A0/A1 for the reported universe (6 at A2) and identically at all three levels on the two restrictions, where it is 45 because a completely measured block leaves nothing undetermined and not because the evidence is stronger. The margins are in win-rate units; the reported-universe margin is the reported leader's *worst case* over completions and is negative, while the two restricted margins are exact differences on a point.

Pattern	target rate	realized coverage	established	reversed	possible	identified
mcar	10%	89.8%	20.0	0	3.4	0
	20%	79.9%	2.4	0	9.8	0
	30%	70.1%	0.0	0	21.6	0
budget	10%	79.1%	13.8	0	8.5	0
	20%	77.9%	13.1	0	8.6	0
	30%	70.0%	5.3	0	12.7	0
block	10%	74.5%	6.2	0	10.4	0
	20%	74.5%	6.2	0	10.4	0
	30%	67.9%	0.5	0	19.8	0
column	10%	83.3%	9.7	0	6.7	0
	20%	78.8%	3.4	0	8.6	0
	30%	69.7%	0.0	0	21.6	0

Table 4: Identification under controlled masking of a completely measured real block, means over 10 seeds, at A1, under HELM's declared metric direction (Appendix F). "Established" and "reversed" count the 45 pairwise orderings inside the true top ten; "possible" counts models that could lead in some completion; "identified" counts models that lead in every completion. Reversed is zero in all 280 analyses, as is identified. At target rates of 50% and above every pattern gives 0.0 established and 31.5–32.0 possible leaders; at 40% the patterns still differ, giving 0.0–0.8 established and 21.5–30.9 possible leaders. The two contrasts §4.2 draws are the mcar-20%/budget-20% pair, matched to within 2 points of realized coverage, and the block-10%/mcar-30% bracket.

Leader query, across 2,400 rows of a 350-configuration grid	rows	share
excluded by the screen (provably not possible)	497	20.7%
admitted to the exact solve	1,903	79.3%
decided	1,811	
time-limited, no verdict	92	4.8% of admitted
decided, a witness is certified (possible)	1,798	
decided, no completion exists (strict gap)	13	0.72% of decided

Table 5: Every model of every configuration put to the cheap per-rival screen, and every model the screen does not exclude put to the exact solve. The screen excludes only rows it can prove impossible (Proposition 3), so the 497 need no solve. The 13 strict gaps are the finding: rows the cheap test certifies as possibly leading for which no completion of the evidence makes the claim true. Their count is a lower bound, since a time-limited row could add to it and cannot remove from it, and the rate is conditional on the solver having decided. Coverage here is a design variable rather than a control: realized coverage runs from 16% to 89% across the five rate settings, while the cell means are nearly equal (48–51%) because every rate is sampled at every universe size (Table 6). The gaps are therefore reported as incidences over the grid, not as a function of coverage.

Level	size	coverage	decided	certified possible	gaps	timeouts
A1	5	49%	207	206	1	0
	6	51%	250	248	2	0
	8	50%	338	338	0	0
	10	48%	337	336	1	92
A2	5	49%	165	162	3	0
	6	51%	210	205	5	0
	8	50%	304	303	1	0

Table 6: Incidence of the sign flip over 350 configurations: $4 \times 5 \times 10 = 200$ at A1 over sizes $\{5, 6, 8, 10\}$ and $3 \times 5 \times 10 = 150$ at A2 over sizes $\{5, 6, 8\}$, at five missingness rates and ten seeds each; realized coverage averaged in the third column. The grid is not a full cross product — no $n = 12$ cell was run at either level (Appendix I) — so the two levels' incidences are not comparable row for row, and the size column is a description of what was sampled rather than a factor whose effect is being estimated. "Decided" counts rows the cheap per-rival test names as possible leaders *and* that the exact solve decided — the only population on which the test's verdict can be checked, since a row that hits its budget returns no verdict to check — and equals "certified possible" plus "gaps" in every row. Rows the test excludes are provably not joint-possible by Proposition 3 and need no solve. All thirteen gaps rest on one route: the exact solve returned `optimal` with a negative joint value, and a separate certificate pass — a feasibility search for a completion, re-scored under the declared rule rather than trusting the solver's objective — agrees, returning `no witness` on all thirteen and on none of the 1,798 possible rows.

Level	$\|M\|$	queried model	cheap bound	exact	slack
A1	5	TNLG v2 (6.7B)	$+0.0000$	-0.0104	0.0104
	6	Cohere medium v20220720 (6.1B)	$+0.0136$	-0.0136	0.0273
	6	BLOOM (176B)	$+0.0091$	-0.0091	0.0182
	10	Cohere small v20220720 (410M)	$+0.0076$	-0.0227	0.0303
A2	5	OPT (175B)	$+0.0455$	-0.0114	0.0568
	5	Jurassic-2 Large (7.5B)	$+0.0000$	-0.0238	0.0238
	5	GPT-NeoX (20B)	$+0.0000$	-0.0208	0.0208
	6	Cohere Command beta (6.1B)	$+0.0182$	-0.0045	0.0227
	6	ada (350M)	$+0.0300$	-0.0400	0.0700
	6	Cohere small v20220720 (410M)	$+0.0000$	-0.0357	0.0357
	6	curie (6.7B)	$+0.0000$	-0.0133	0.0133
	6	J1-Jumbo v1 (178B)	$+0.0000$	-0.0077	0.0077
	8	ada (350M)	$+0.0040$	-0.0198	0.0238

Table 7: Every row on which the cheap per-rival test is wrong, out of 1,811 decided rows. "Cheap bound" is $\min_j \sup_F(\rho(r) - \rho(j))$ and "exact" is $\max_{S' \in F} \min_j(\rho(r) - \rho(j))$; a sign flip between the two columns is the gap. All thirteen are real configurations of the completely measured 32×22 block under the block mask, with the seeds and rates recorded in `exp_c_joint_gap/results.jsonl`. **Six of the thirteen have a cheap bound of exactly** $+0.0000$ — the boundary case of §4.3, where the relaxation's best completion leaves r tied with a rival rather than ahead of it; a rule admitting a row only on a *strictly* positive cheap bound would therefore find seven gaps rather than thirteen. The universe sizes are those the grid sampled and are not a statement about where gaps do and do not exist: the largest is the largest the exact solve could still be trusted on.

Realized coverage	rows	decided	decided %	gaps / decided	gap rate
$[0.00, 0.35)$	784	644	82%	$7/644$	1.09%
$[0.35, 0.55)$	636	555	87%	$1/555$	0.18%
$[0.55, 0.70)$	428	308	72%	$1/308$	0.32%
$[0.70, 1.01)$	552	262	47%	$3/262$	1.15%

Table 8: Where the gaps sit by realized coverage, reported as counts and fractions. The bands are nested subsets of one design and the rows are dependent, so no interval is given and no ordering is read from the column.

	b_1	b_2	b_3
m_1	3	—	2
m_2	—	1	**7**
m_3	—	9	—
m_4	9	10	2
A_b under A1	$[3, 9]$	$[1, 10]$	$[2, 2]$

Table 9: A measured instance in which a dynamic A1 update turns a refuted claim undetermined. Column b_3 has two observed values, both 2, so its A1 range — the observed span of that column, not a platform declaration — is the degenerate $[2, 2]$. Measuring m_2 in that column returns **7**, outside it. Under the frozen contract the slice $F(q{=}7)$ is *empty*: the contract, not the analysis, has been contradicted, and the only consistent response is to widen the range to $[2, 7]$ and restart the chain. Doing so frees m_3 to take values it could not take before, and the claim "m_3 beats m_4" goes from refuted, $[-0.500, -0.111]$, to undetermined, $[-0.500, +0.111]$. The undetermined count can rise, so re-deriving the range from the data after every measurement — a *dynamic* A1 — has no certificate to offer at all. This is why the episode freezes its ranges and fixes only the acquired cells.

	probe	state	candidate scores and choices that moved
deployable	uniform	identical	0 of 426 scores, 0 of 54 picks
deployable	out of range	identical	0 of 426 scores, 0 of 54 picks
deployable	constant	identical	0 of 426 scores, 0 of 54 picks
control	hindsight oracle, candidate keys moved		486 of 1,434 (34%)
control	`acquire` at the probe cell		state changes, every probe

Table 10: The outcome-independence audit. The deployable objectives are scored on 426 state–probe evaluations — every free cell of one 79.8%-observed instance under three probes — and the choices of all six deployable rules are compared on three instances, three probes each: the four objectives of the main family, plus the two tie-break controls of §J, which read the same state and differ only in how they break a tie. Not one number they read moves when every withheld value is replaced, while the control, the outcome-aware arm, moves on a third of the candidate keys under the same probes: a leak in a deployable arm had a third of the candidate cells to show up in and appeared in none. An argmax-only control was tried first and is *inert* on two of the three instances — the outcome-aware arm's keys tie at $(0, 0)$ and its argmax is then fixed by iteration order — which is why the control is measured on the whole key vector.

Rule	mcar	budget	block	all 12	worst run
gain-contract	25.25	19.25	15.00	19.83	10
contract-gain	25.25	18.00	**15.50**	19.58	10
expgain	**28.00**	**21.75**	15.00	**21.58**	**14**
gain-expgain	27.25	**21.75**	15.00	21.33	**14**
gain-rand	17.75	14.25	12.50	14.83	2
gain-widest	23.75	16.75	12.50	17.67	5
one-step greedy oracle (reads the outcome)	28.25	20.50	16.50	21.75	13
widest	23.00	17.00	13.50	17.83	6
complete	19.50	17.00	10.00	15.50	6
random	14.00	14.50	10.50	13.00	2

Table 11: Of the 45 comparisons in the reported top ten, how many are established after 20 additional measurements: mean over the twelve runs, grouped by pattern. Realized coverage is 78.7–90.1% (mcar), 74.1–83.7% (budget) and 72.4–78.1% (block); the four mcar runs are distinct instances, the four budget runs contain one coincident pair, and the four block runs are two distinct masks each appearing at both target rates, so the twelve runs are nine distinct instances. The first six rows and the last three read only the published matrix; the outcome-aware row reads the value it is about to acquire. Bold marks the best deployable value in each column, except on that row, which is shown unbolded as a comparator rather than a competitor. "Worst run" is the minimum over the twelve.

	steps decided by (of 240)					realised	
Rule (shared primary key G)	class 1	class 2	class 3	class 4	$\sum_t G$	drop	mean /45
`gain-rand` (G, rand)	28	12	16	212	29	63	14.83
`gain-widest` (G, widest)	37	17	20	203	37	99	17.67
`gain-contract` (G, C), proposed	42	18	24	198	42	123	19.83
`gain-expgain` $(G, \bar{G})$	21	11	10	219	21	142	21.33

Table 12: The attribution experiment. All four rules have the same primary key, so every difference between the rows is the tie-break. Class 1 is "some candidate has $G > 0$" and is the union of classes 2 and 3; class 2 is "G has a unique maximiser"; class 3 is "the maximisers tie and the secondary key decides"; class 4 is "$\max_q G(q) = 0$ over every candidate". "Drop" is the total number of target orderings the twelve runs settled between them, out of $12 \times 45 = 540$, and $\sum_t G$ is the certified gain over the same runs; the last column is the mean final count.

Work	Unknown object	Query / action	Objective	Answer model assumed
Pini et al. (2007)	an incomplete preference profile (given, not chosen)	— none: complexity of a given profile	is an alternative a necessary / possible winner	no
Lu & Boutilier (2020)†	voters' partial rankings	a preference query to a voter	minimise minimax regret of the chosen outcome	no: robust optimisation over the given partial rankings
Shah & Yu (2026)	agents' full preference rankings	a next-best query (agent names its next favourite)	competitive ratio against hindsight query count for calling the winner	no
Mohankumar & Khapra (2022)	pairwise preference probability between two systems	one pairwise comparison of two systems	identify the top-ranked system	**yes**: BTL, BTL-logistic or linear, with Bayesian uncertainty
Li et al. (2025)	a model's unobserved prompt-level scores	run one more prompt for one model	error of the estimated benchmark scores at a fixed prompt budget	**yes**: a neural process, i.e. a latent-variable generative model over scores
Arviv et al. (2026)	a model's true mean score	more items, then stop or continue	stopping time at a target error level	**yes**: a sampling distribution for item scores
this work	a benchmark score in a declared interval	one cell $(\mathrm{model}, \mathrm{benchmark})$ of the published matrix	published comparison claims the evidence comes to settle	no: worst case over the admissible box

Table 13: What each line's query buys, and what it assumes about the answer. The six lines above the rule are elicitation work in the preference and evaluation literatures; the last is the acquisition of Appendix J. The distinction the table is drawn to show is in the last column: the three evaluation-specific lines model the answer distribution, the three preference lines do not, and neither group's unknown object is an interval-valued benchmark score. †Read from the publisher's abstract only; the version of record is not openly available. No cell in this row is populated from anything the abstract does not state, and the abstract states the partial-ranking input, the minimax-regret objective for the chosen outcome, the preference queries, and the tractability split.

Work	Guarantee	Computational form	Targets an already-published claim?	Interval scores and set identification
Pini et al. (2007)	complexity classification (P / NP-hard) for winner determination	combinatorial, over the profile as given	no	no
Lu & Boutilier (2020)†	minimax-regret bound on the chosen outcome	tractable for some rules, intractable for others	no	no
Shah & Yu (2026)	worst-case competitive ratio; $O(\sqrt{m})$ for Borda	level-wise pruning; multi-scale score thresholding	no	no
Mohankumar & Khapra (2022)	regret bounds under stochastic feedback	dueling bandits with upper confidence bounds	no	no
Li et al. (2025)	none: empirical comparison of policies	an RL policy over prompt subsets, with entropy and mutual-information baselines	no	no
Arviv et al. (2026)	Type-I error control across interim analyses	sequential tests (Pocock, SPRT)	no: it decides when a running evaluation has enough evidence	no
this work	worst-case lower bound on settled claims, held *before* the query, under a named range contract (Proposition 7)	closed form, $O(NB\,\lvert Y_b\rvert\,P)$, no solver and no outcome model	**yes**: the target set is the orderings the platform already published	**yes**: an identified interval per claim, labelled by set membership

Table 14: What each line guarantees and what it identifies. "Targets an already-published claim" asks whether the query is chosen to resolve a comparison the platform has *already reported*, rather than to call a winner or to stop; "interval scores and set identification" asks whether an individual score is an interval and a claim's status is membership of a set of completions. The perimeter this paper claims is the joint occurrence of the two **yes** cells, not the idea of scoring a candidate query, which is not new here. The Pini et al. row is the IJCAI 2007 conference paper, which is what was read; the journal version is paywalled and contributes no cell. †As in Table 13: abstract only, and no cell beyond what it states.