WHEN DO HARDER TASKS PRESERVE AGENT RANKINGS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Adding harder tasks is a natural way to refresh an agent benchmark, but it can reduce rather than improve resolution when strong systems fail those tasks together. We introduce **measurement-regime diagnosis**, a post-pilot audit that tests whether hardness agrees with finite-budget ranking recovery and frontier-local distinguishability. On SWE-bench Verified, the 50 hardest and frontier-informative tasks are disjoint, and 96% of the hard set is all-fail among frontier systems. On Terminal-Bench 2, the same diagnostics largely retain hardness as a candidate, although the alignment is outcome-encoding sensitive. SWE-bench transfer changes ordering with budget. Highest frontier information (HFI) has higher fixed-reference correlation than hardness at 25 and 50 tasks, while hardness catches up at 100. None improves the anchor. The pattern survives 10 task splits, repository partitions, and budgets matched on pre-cutoff estimated API calls. At $k = 100$, ranges cross zero, and one family omission changes the $k = 50$ ordering. Across four Terminal-Bench 2 chronology/missingness variants, HFI-minus-hardest recovery is positive, but the best policy changes. We therefore formulate a regime-, target-, and budget-indexed Pareto decision with abstention. The evidence is retrospective and conditional on observed systems and piloted candidate outcomes. It does not establish zero-shot utility for newly authored tasks or an absolute benchmark ceiling.

1 INTRODUCTION

Agent benchmarks evolve as tasks are added, retired, or hardened. A common refresh rule adds the hardest available tasks, but hardness helps ranking only while preserving system distinctions. A task that every frontier system fails lowers scores without ordering that panel.

We ask: **after candidate tasks have been piloted, when is hardness a defensible maintenance policy?** The answer depends on target systems, task budget, global versus frontier-local comparison, and non-tied pair support.

We introduce **measurement-regime diagnosis**, a lightweight audit of an observed agent-by-task matrix. It compares the hardest set with an empirical frontier-information set, measures saturation, evaluates finite-budget ranking recovery, checks row dependence, and abstains when effective comparisons are too sparse. This is a rank-resolution diagnostic, not a complete account of benchmark value. Hardness remains a useful but regime-dependent proxy.

The empirical picture is budget-dependent ordering, not a universal winner. SWE-bench's hardest and frontier-informative Top-50 sets are disjoint. In frozen transfer HFI leads at 25 and 50 tasks, hardness catches up at 100, and no policy improves the strong anchor overall. Terminal-Bench supplies a conditional, encoding-sensitive aligned contrast.

In short, 48 of the 50 hardest SWE-bench Verified tasks are all-fail among the observed frontier systems. HFI has higher global fixed-reference correlation at $k = 25/50$, whereas hardest is competitive at $k = 100$. These results do not identify one policy as a universal winner.

We make three contributions:

1. **A maintainer-facing screening and abstention tool, not a selection oracle.** We turn difficulty, task-set overlap, frontier saturation, effective-pair coverage, and row dependence

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.

into a four-step decision protocol with an explicit abstention rule. The tool rules out unsupported choices and presents a Pareto set rather than naming one point-decision winner.

2. **A budget-dependent empirical finding.** Frozen SWE-bench transfer shows that the descriptive policy ordering changes with task budget and evaluation target. HFI's global fixed-reference advantage at $k = 25/50$ is positive in all 10 deterministic splits, whereas the $k = 100$ range spans zero. Thus, "pick harder tasks" and "pick informative tasks" are both incomplete prescriptions.
3. **Robustness beyond one split or proxy.** We test 10 deterministic and within-repository partitions, budgets matched on pre-cutoff estimated API calls, stronger compression comparators, calibration regret, family/repository deletions, and Terminal-Bench encoding, chronology, and missingness variants. These are correlated sensitivities, not independent replications.

2 Problem Setting

Let $Y \in \{0, 1, \mathrm{NaN}\}^{|A|\times|T|}$ contain system-by-task outcomes. Rows may be models, submissions, scaffolds, or conditions. They are neither independent nor flawless capability ground truth. We use *agent ranking* as shorthand for ordering these evaluated system configurations. The matrix is the measurement instrument under study.

For row success s_a and quantile q, define the observed frontier $F_q = \{a : s_a \geq \mathrm{quantile}_q(s)\}$. This is a target population, not an absolute capability threshold. Static analyses also cannot identify it online before any outcomes are observed.

For observed rows O_t, task success is $p_t = \mathrm{mean}_{a\in O_t} Y_{a,t}$ and difficulty is $d_t = -\mathrm{logit}(\mathrm{clip}(p_t))$. Let r_t denote the point-biserial association between task outcome and standardized logit row success, not a fitted item parameter. We also record frontier success $p_{t,F}$, variance $p_{t,F}(1 - p_{t,F})$, saturation, and a leave-one-item-out ablation.

Our frontier-information score is descriptive and item-information-inspired. Set $\alpha_t = \max(r_t, 0)$ and $\beta_t = d_t$ as a difficulty-derived intercept offset, not a conventional 2PL location. For standardized row scores θ, form a padded frontier-tail grid of 201 points from $\min(-3, \min(\theta) - 0.5)$ to $\max(3, \max(\theta) + 0.5)$, restrict it at the observed frontier quantile, and average $\alpha_t^2 P_t(\theta)(1 - P_t(\theta))$, where $P_t(\theta) = \sigma(\alpha_t\theta - \beta_t)$. The largest scores define highest frontier information (HFI). Hybrids add frontier variance or local association.

These lightweight IRT-inspired proxies are not fitted item information, a latent-trait model, or a reliability estimate. They compare frozen policies in an observed post-pilot matrix. Their target is conditional rank-resolution value.

In **dated row transfer**, utilities use systems recorded by cutoff c. Selections are frozen and evaluated on later rows against the unselected-task complement.

In **release-defined task-pool transfer**, the Lite/Verified intersection is anchor T_0 and Verified-only tasks form T_{new}. A pre-analysis hash partitions candidates C and fixed reference R. Utilities use pre-cutoff outcomes on C, while later systems are scored against untouched R. Candidate outcomes are observed, so this is post-pilot, not zero-shot.

2.1 Evaluation targets

No single metric captures benchmark maintenance. We report:

- all-system Spearman correlation with a held-out or fixed-reference ranking.
- frontier-only Spearman and Kendall correlations.
- pairwise accuracy conditional on pairs that are non-tied under both selected and reference scores.
- effective-pair count and non-tied pair coverage.
- frontier all-fail, all-success, and low-variance rates.
- task-set overlap relative to the expectation for random size-k sets.

A policy can recover the global ranking while tying most frontier pairs. Low variance means frontier success at most 0.05 or at least 0.95. all-fail/all-success means `0/1`. Primary budgets count tasks. One labeled sensitivity instead matches pre-cutoff expected reported API calls, an interaction-count proxy rather than runtime, tokens, FLOPs, or exact future cost.

3 MEASUREMENT-REGIME DIAGNOSIS

The audit produces a bounded diagnosis, not a universal task score. It is indexed by outcome matrix, target system population, frontier quantile, task budget, utility, and outcome encoding.

Audit protocol

1. **Estimate.** Compute task success, difficulty, empirical row-task association, frontier success and variance, saturation, and available information scores.
2. **Diagnose.** For hardest set H_k and utility set I_k, report difficulty-information association, overlap $o = |H_k \cap I_k|/k$, random-set expectation k/N, chance-adjusted overlap, and frontier saturation. A hard tail that is both separated from I_k and all-fail at the frontier is a measurement mismatch.
3. **Validate.** Freeze each selected set before evaluation rows or reference tasks are scored. Compare global ranking recovery, frontier-local metrics, saturation, split sensitivity, and dependent-row variants. Trials are Monte Carlo partition repetitions, not independent benchmark replications. Unless stated otherwise, percentile ranges and correlated-cell counts describe resampling or partition sensitivity, not IID confidence intervals or independent-sample counts.
4. **Nominate, veto, or abstain.** Ranking diagnostics nominate a Pareto set. An independently authored construct blueprint and constraints on domain, safety, security, licensing, and accessibility may veto any nominee. Select only for the stated target and budget, and abstain when row diversity or effective-pair support is insufficient. Section 11 gives a prospective tau2-airline instance in which this rule stops selection before a regime statistic is computed.

This two-layer rule is substantive: an all-fail task is uninformative for current-panel rank resolution, not necessarily scientifically useless. A separate reserve of hard, novel, or safety-critical tasks should be maintained independently of current outcome-derived utility.

For the static SWE-bench grid, we use task splits `40/40/20` and `30/50/20`, calibration-row fractions `0.5/0.7`, frontier quantiles `0.7/0.8/0.9`, budgets `10/25/50/100`, and 100 trials. Each trial re-estimates candidate properties on calibration rows using initial-benchmark scores as the empirical row-performance score, then evaluates selected tasks on disjoint evaluation rows and held-out task columns.

The output is a decision card with regime, target, declared task or resource budget, dependence warning, saturation, Pareto candidates, and claim boundary. A **difficulty-information mismatch** combines below-chance hard/utility overlap with hard-tail saturation. A **strong aligned regime** has above-chance overlap and low saturation. A **weak aligned / saturated / row-limited contrast** has thin or dependent rows or substantial saturation. Terminal-Bench 2 is reported more narrowly as a **conditional aligned contrast** because its conclusion is encoding-sensitive.

4 BENCHMARK MATRICES AND REGIME SUMMARY

SWE-bench Verified contains 81 public leaderboard rows and 500 tasks, with no missing cells. Forty rows duplicate another complete outcome vector, leaving 41 representatives for dependence and temporal analyses. Terminal-Bench 2 has 75 rows and a 90-task union from archived results. Most systems have 89 observed cells. The row-limited tau2_telecom matrix has 10 dependent model/condition rows and 114 tasks. tau2_telecom is appendix-level evidence because of row count and saturation.

Cell semantics differ. SWE-bench uses binary resolution. Terminal-Bench 2 thresholds mean valid verifier reward at 0.5 in the primary analysis and treats no valid reward as missing. tau2 thresholds a four-trial mean at 0.5. We compare regime patterns, not absolute score levels.

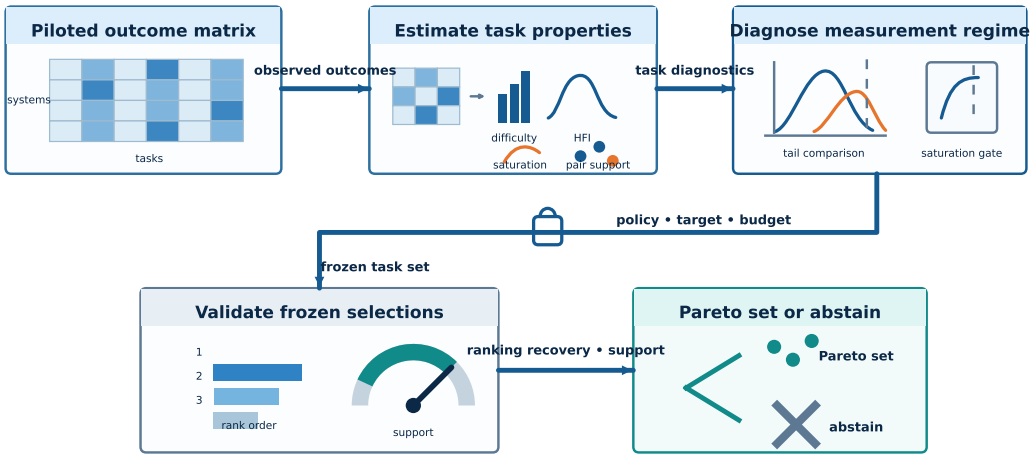

Figure 1: Measurement-regime diagnosis maps a piloted outcome matrix to a budget-specific maintenance decision. For SWE-bench, Top-50 hardest/HFI overlap is 0/50 and 48/50 hardest tasks are all-fail among observed frontier systems. Sets are frozen before validation. At k=25/50, HFI has higher global fixed-reference correlation; at k=100, the split range spans zero. The output is a Pareto set or abstention, not a selection oracle.

Table 1: Regime diagnostics across benchmark matrices.

Benchmark	Rows/tasks	k/N	Adjusted overlap	Hard all-fail	Diagnosis
SWE-bench Verified	81 x 500	0.10	-0.111	0.960	mismatch
Terminal-Bench 2	75 x 90	0.56	0.595	0.020	conditional aligned

All rows use Top-50 sets; k/N exposes the unequal selected fraction. Chance-corrected overlap is $(o - k/N)/(1 - k/N)$. Terminal-Bench uses mean reward threshold 0.5; Fig. 3 reports encoding sensitivity. tau2 is appendix-level.

The regime-diagnostics table summarizes the two primary matrices, their outcome semantics, alignment, Top-k overlap, frontier saturation, and regime label. The tau2 row-limited contrast and auxiliary execution gate remain in the appendix because they cannot support a third independent regime claim.

5 SWE-bench Verified: A Mismatch Regime

5.1 Static diagnosis

SWE-bench Verified is the clearest mismatch case. Difficulty-information Spearman is -0.1050, with task- and row-resampling intervals that include zero. The stronger evidence is finite-budget separation: Top-50 hardest/HFI overlap is 0.0000 versus random expectation 0.1000 (chance-adjusted -0.1111), while hard-tail frontier all-fail is 0.9600. Collapsing duplicate rows, alternative proxies, quantiles, resampling, and family/repository omissions retain this separation-plus-saturation diagnosis. The all-fail row-resample and row-subset ranges are both [0.9400, 1.0000].

5.2 Frozen dated-row transfer

Three cutoffs yield `15/26`, `21/20`, and `28/13` calibration/future rows. Over 27 correlated cells (three cutoffs, three frontier quantiles, and three budgets), HFI/hardest mean future complement Spearman is `0.8093/0.5561`, but only `7/27` cells retain 10 effective pairs per strategy. no frontier-local dominance is supported.

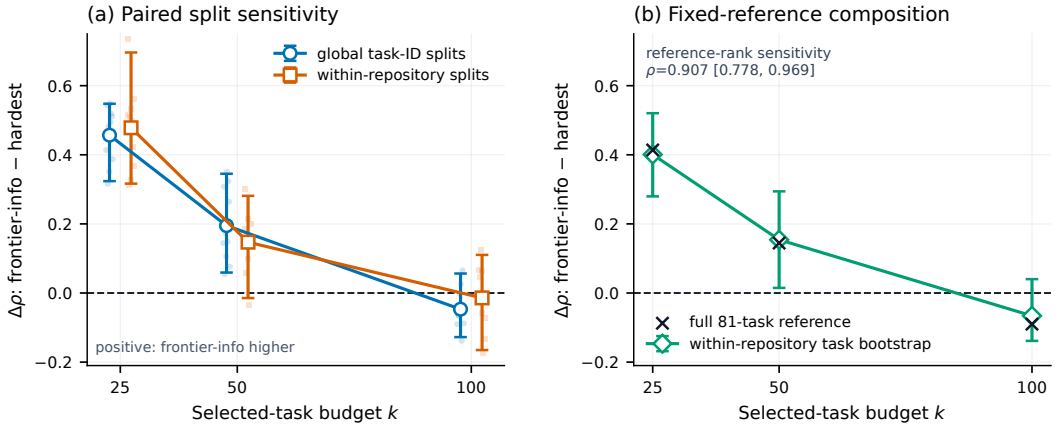

Figure 2: Budget-dependent descriptive ordering in release-defined SWE-bench transfer. HFI has higher global fixed-reference correlation than hardest at k=25/50, whereas the k=100 split range spans zero. HFI is the empirical heuristic defined in Section 2. Panel (a) shows paired differences across 10 deterministic task splits; panel (b) shows 2,000 within-repository reference-task bootstrap draws with frozen selections. Bars are 2.5–97.5 percentile sensitivity ranges, not IID confidence intervals. The budget interaction is stable, while k=100 ranges cross zero.

5.3 Release-defined new-task-pool transfer

A pre-analysis split assigns 407 Verified-only tasks to 326 candidates and an 81-task fixed reference; 93 shared tasks form the anchor. Across 27 settings, HFI/hardest fixed-reference Spearman is `0.7349/0.5788`, changing from `0.7433/0.3294` at $k = 25$ to `0.7089/0.7989` at $k = 100$. The anchor reaches 0.8423, and no strategy improves it on average.

Across 10 splits, HFI-minus-hardest at $k = 25/50/100$ is `0.457 [0.324, 0.548]`, `0.195 [0.059, 0.345]`, and `-0.047 [-0.128, 0.056]`; HFI leads in `10/10`, `10/10`, and `2/10`. Repository sensitivity retains the low-versus-high interaction, whereas the $k = 100$ range spans zero. S12 pools a different set of cells and is not directly comparable.

The API-call-matched sensitivity uses 41 complete, outcome-unique `bash-only` rows. HFI-minus-hardest fixed-reference Spearman averages 0.1373 over 27 correlated cells and is `0.3883/0.1634/-0.1399` at $k = 25/50/100$; realized future calls average 1.3498 times the frozen budget. Only `3/27` cells pass the strict 10-pair gate. Exact common-pair comparisons favor hardest, so no frontier-local advantage is claimed.

5.4 Split, comparator, and policy sensitivity

Across 10 deterministic partitions, HFI leads at $k = 25/50$. At $k = 100$, hardest and Hybrid(0.5) reach about 0.799 versus 0.752 for HFI. These are correlated sensitivity checks, not independent replications.

Repository-stratified splits and compression comparators retain the low-budget ordering. The sign pattern remains in `9/10` leave-one-family-out variants, but only `6/10` preserve every identifiable cutoff; removing OpenAI changes HFI-minus-hardest to `0.191/-0.052/-0.172` at $k = 25/50/100$. Calibration-only policy choice matches the hindsight oracle in only `12%` to `38%` of cells, so the audit retains a Pareto set rather than automatically naming a winner.

6 Objective Tradeoffs, Not a Universal Winner

One declared scalarization, `hybrid_global_frontier_lambda_0.5` (Hybrid(0.5)), filters for coverage at least 0.25, low variance at most 0.65, and all-agent Spearman within 0.08 of the best, then combines global and frontier metrics with fixed weights. It selects Hybrid(0.5), with Spearman

Table 2: Release-defined transfer across 10 deterministic candidate/reference task splits. Cells show mean future fixed-reference Spearman and 2.5–97.5 percentile split-seed sensitivity ranges, not confidence intervals. HFI denotes highest frontier information.

Selection policy	$k = 25$	$k = 50$	$k = 100$
Hardest	0.251 [0.123, 0.377]	0.589 [0.478, 0.719]	0.799 [0.739, 0.866]
HFI	0.708 [0.644, 0.778]	0.784 [0.713, 0.832]	0.752 [0.697, 0.816]
Max entropy	0.718 [0.654, 0.768]	0.747 [0.665, 0.794]	0.773 [0.663, 0.840]
Hybrid (0.5)	0.635 [0.583, 0.716]	0.721 [0.664, 0.766]	0.799 [0.728, 0.861]
Greedy reconstruction	0.689 [0.598, 0.849]	0.722 [0.608, 0.806]	0.790 [0.676, 0.860]
Diversity + entropy	0.688 [0.549, 0.777]	0.754 [0.658, 0.798]	0.782 [0.699, 0.846]
Δ(HFI−hardest), global	+0.457 [+0.324, +0.548]	+0.195 [+0.059, +0.345]	-0.047 [-0.128, +0.056]
Δ(HFI−hardest), within-repo	+0.478 [+0.317, +0.696]	+0.147 [-0.015, +0.281]	-0.014 [-0.165, +0.110]

Candidate outcomes are available only on calibration rows; reference tasks and future rows are excluded from selection. Brackets are deterministic partition-sensitivity ranges, not confidence intervals. Paired rows first average the shared 3 cutoffs $\times$ 3 frontier quantiles within each split seed. Positive Δ favors HFI.

0.8962, effective-pair accuracy 0.5859, coverage 0.7381, and low variance 0.3925. Thresholds and weights are maintainer preferences, not discovered constants.

The supplementary table shows the same geometry. HFI has Spearman 0.9158, coverage 0.6884, and low variance 0.6378. Max entropy has Spearman 0.8972, pair accuracy 0.5724, coverage 0.7084, and low variance 0.4781. Hardest obtains 0.6229, 0.5733, 0.5147, and 0.7527. Random obtains 0.8454, 0.5586, 0.6464, and 0.7306. All strategies use the same pre-specified 100-trial grid.

Across 48 grid settings, Hybrid(0.5), unconstrained hybrids, and constrained hybrids win `8/48`, `21/48`, and `16/48`. This supports a Pareto family more than one lambda. In transfer, HFI leads limited-budget global recovery, entropy can lead supported local comparisons, hybrids can improve coverage, and hardness recovers at broad budget. Either tail can saturate.

SWE-bench maintenance decision card

- **Target and budget:** specify all-system ranking, frontier-local ordering, or both, with $q \in \{0.7, 0.8, 0.9\}$ and a concrete task budget.
- **Diagnosis:** hardest/HFI separation and hard-tail saturation survive exact duplicate collapse. The budget interaction is stable across task splits but the $k = 50$ ordering is not invariant to every family omission.
- **Pareto candidates:** HFI for limited-budget global recovery, entropy or anti-saturation for frontier variation, Hybrid(0.5) for a declared tradeoff, hardness only where a broad budget or local ordering supports it.
- **Decision or abstention:** retain multiple policies when calibration regret is high. Abstain with fewer than 10 effective pairs or a narrow system panel.
- **Boundary:** this is post-pilot selection from observed candidate outcomes, not zero-shot scoring of newly authored tasks.

7 Terminal-Bench 2: A Conditional Aligned Contrast

Terminal-Bench 2 prevents an anti-hardness interpretation, but the diagnosis remains budget-indexed. Under the mean-reward threshold 0.5, difficulty-information Spearman is 0.5506. Hardest/HFI overlap at $k = 10/20/30/50$ is `0.000/0.400/0.633/0.8200`. Thus Top-10 diverges even though Top-50 is conditionally aligned. At $k = 50$, the hardest all-fail rate is 0.0200. Under independent uniform Top-50 sets, expected overlap is 0.5556; chance-adjusted observed overlap is 0.5950 (one-sided combinatorial exact $p = 1.22 \times 10^{-8}$). This null calibrates overlap, not population evidence.

The row-bootstrap interval for Spearman is [0.3614, 0.6697], its row-subset range is [0.3948, 0.5898], and Top-50 overlap spans [0.7400, 0.8400]. Leave-one-item-out retains 0.980 of the HFI Top-50.

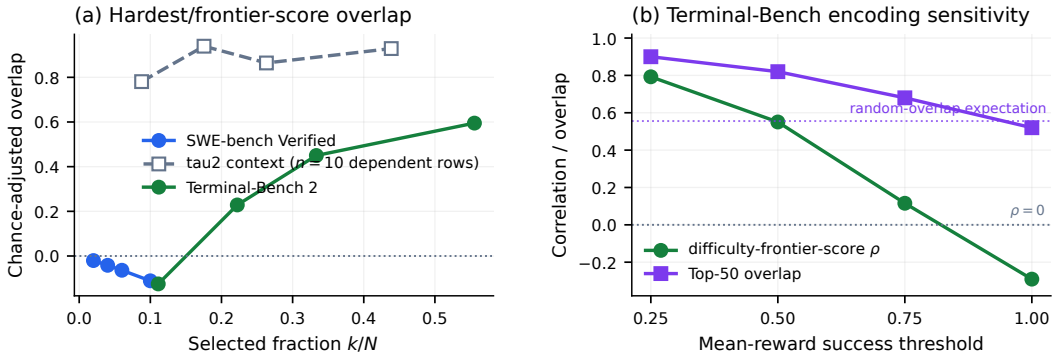

Figure 3: Conditional alignment across benchmark regimes. Panel (a) reports chance-corrected overlap (o-k/N)/(1-k/N) and plots the selected fraction k/N; tau2 is shown only as gray row-limited context. Panel (b) shows that the Terminal-Bench diagnosis varies with reward threshold even though hard-tail all-fail remains low. The tau2 curve is an appendix-level row-limited contrast.

Overlap at $q = 0.7/0.8/0.9$ is `0.800/0.820/0.820`. The primary aligned diagnosis is therefore not driven by one system subset.

It is, however, outcome-encoding sensitive. At reward thresholds `0.25/0.50/0.75/1.00`, Spearman is `0.793/0.551/0.115/−0.291` and Top-50 overlap is `0.900/0.820/0.680/0.520`. A continuous mean-reward proxy gives Spearman 0.579 and overlap 0.820. Hard-tail all-fail remains low throughout, between 0.000 and 0.060, so the stable claim is that the hard tail does not collapse into universal frontier failure. The strength of difficulty-information alignment is not encoding-invariant.

Retrospective chronological transfer admits 62 of 75 rows and excludes 13 timezone-naive rows. HFI-minus-hardest selected-versus-complement Spearman averages `0.0959/0.0497` under available-case/complete-task handling and `0.0793/0.0418` under two analysis-mirror variants. It is positive in `24/17/24/21` of 27 correlated cells, but the highest-mean policy changes across variants: Hybrid(0.5), anti-saturation, random, and random. Strict pair support covers only `10/27`, `7/27`, `10/27`, and `4/27` cells. This supports the Pareto rule, not a universal HFI win. Neither ordering is publication chronology.

The cross-benchmark overlap figure compares chance-adjusted overlap by selected fraction and shows this threshold sensitivity. The audit retains hardness as a Pareto candidate only where the declared encoding and budget support it; it does not establish that hardness itself recovers the reference ranking.

8 Related Work

8.1 Efficient benchmark subsets

Efficient benchmark subsets use IRT, ranking stability, adaptive sampling, or learned construction (Perlitz et al., 2024) (Maia Polo et al., 2024) (Kipnis et al., 2025) (Saranathan et al., 2025) (Wang et al., 2026). The closest agent studies filter historical pass rates for rank fidelity under shifts (Ndzomga, 2026) or replay completed benchmarks under coverage and unresolved-pair constraints (Huang, 2026). Active Evaluation Acquisition learns queries from cross-item dependence (Li et al., 2025b), while amortized model-based evaluation predicts item difficulty and combines IRT with adaptive item generation (Truong et al., 2025). These estimate efficient scoring. We freeze one public set for dependent systems and audit rank recovery, frontier support, saturation, and policy-level abstention.

BenchScope is the closest matrix-audit neighbor: it uses population-conditional effective dimensionality and spectrally diverse selection, including a SWE-bench design/future split (Sha & Zhao, 2026). We ask a narrower question—whether hardness supports fixed-budget global and frontier-local decisions—and compare transparent maintenance policies. We do not claim the first population-conditional matrix diagnostic, maintainer workflow, subset optimizer, or IRT estimator.

8.2 Items as measurements

Item-informativeness and item-response analyses show that examples differ in difficulty, discrimination, and value for comparing systems (Rodriguez et al., 2021) (Vania et al., 2021). Adaptive and fluid evaluation choose items as a function of estimated ability (Hofmann et al., 2025) (Li et al., 2025a). Recent work also questions whether standard IRT assumptions are reliable for AI evaluation (Jiang et al., 2026). We use psychometric quantities as transparent diagnostics, but public fixed-form maintenance must choose one task set for a population of dependent systems.

8.3 Benchmark lifecycle and agent substrates

Benchmark saturation, rank uncertainty, dynamic updates, and documentation motivate lifecycle-aware evaluation (Akhtar et al., 2026) (Neuhof & Benjamini, 2026) (White et al., 2024) (Sokol et al., 2024) (Reuel et al., 2024). Measurement-regime diagnosis is a narrower operational module: it tests whether "harder" is the right next direction for a declared target and budget.

SWE-bench-Live continuously curates fresh repository issues and executable environments, and shows that recency is not itself a hardness proxy (Zhang et al., 2025). Our release-defined transfer neither constructs nor executes live tasks: it compares post-pilot retention policies using observed candidate outcomes and an untouched reference set.

AgentBench, tau-bench, SWE-bench, SWE-bench Verified, and Terminal-Bench supply the agent-evaluation setting (Liu et al., 2023) (Yao et al., 2024) (Jimenez et al., 2023) (SWE-bench, 2024) (Merrill et al., 2026). We analyze result matrices as measurement instruments rather than proposing a new benchmark, agent, or runner.

9 Discussion

9.1 What the budget-dependent ordering means

The evidence establishes a descriptive retrospective ordering change, not a population-level reversal. At 25 or 50 tasks, frontier-varying tasks have higher mean fixed-reference rank correlation than the hardest tail. At 100, broad coverage makes hardness competitive. Maintainers should ask "hard for whom, under which budget, and for which ranking target?"

The strong 93-task anchor also matters: no candidate policy improves anchor-plus-selected ranking on average, so retaining the anchor and deferring expansion is a candidate option, not an established optimum.

9.2 Why the audit does not choose automatically

Calibration regret is substantial, and the calibration winner rarely matches the hindsight future winner. The audit therefore screens unsupported choices, presents a Pareto set, and asks maintainers to declare the primary target. It cannot certify construct coverage from response geometry.

9.3 Recommended maintenance workflow

For a piloted pool, predeclare population, encoding, budget, and utility. Report alignment, chance-adjusted overlap, saturation, pair coverage, and dependence. Compare transparent policies, then freeze sets for later rows or an excluded reference. Thin comparisons or split-sensitive policy choices require abstention or a multi-policy pilot.

Selection passes a separate governance gate: a blueprint sets minimum coverage, while safety, security, licensing, accessibility, and stakeholder constraints can veto a rank-efficient set. Maintainers should reserve hard/novel tasks outside incumbent outcomes, rotate sealed panels, and audit family balance and gaming. The decision card records the owner, stakeholders, criterion, evidence, sign-off, veto, and review date. A missing required security stratum triggers rejection or re-optimization. These controls remain unvalidated.

10 CONCLUSION

Harder tasks help when they distinguish the target panel. All-fail items cannot resolve current-panel ranks but may support coverage, safety, or future-capability headroom. The audit rejects SWE-bench's saturated hard tail, conditionally retains Terminal-Bench hardness, and reveals budget-dependent SWE transfer. Benchmark maintenance is a conditional measurement decision, not a monotone search for hardness.

11 LIMITATIONS

All analyses are retrospective and conditional. SWE dates, Terminal completion, and analysis-mirror reconstruction are imperfect chronology proxies. None proves publication chronology or live acquisition. Release membership and the reference are frozen, but candidate outcomes are observed during calibration. The API-call sensitivity covers 41 complete, outcome-unique rows and excludes tokens, latency, FLOPs, and realized cost.

The 93-task Lite/Verified intersection is codebase-imbalanced (Django is about 46%). Repository stratification and reference bootstraps do not remove content dependence. Leaderboard rows may share models, scaffolds, data, harnesses, and practices. Cutoffs, partitions, deletions, and bootstraps are correlated sensitivities; family labels are name-based, not verified lineages.

HFI is a matrix-derived heuristic, not a calibrated IRT quantity. Its scores inherit sampling and dependence limitations, and the padded grid is one design choice. Bounded-range and empirical-score variants change task identities but preserve the regime and budget ordering. Effective-pair accuracy can condition policies on different pairs; the exact-common-pair audit leaves only `19/5/2` cells at gates 5/10/20, supporting abstention rather than a reliability claim.

SWE-bench and Terminal-Bench carry the main contrast. The dependent, row-limited tau2_telecom panel has Spearman 0.9923, Top-50 overlap 0.9600, random expectation 0.4386, chance-adjusted overlap 0.9287, and HFI low variance 1.0000, so it remains appendix-level. Separately, a prospectively pre-specified, outcome-blind generation-6 calibration run on tau2-airline stopped at its predeclared diversity gate before any regime statistic was computed. The tiny10 execution gate is a stopping diagnostic, not positive task-acquisition evidence.

The generation-6 calibration cohort covered four checkpoint pairs across two declared families. Its three outcome-diversity gates failed as specified: unique candidate outcome vectors `2 < 8`, unique anchor scores `2 < 4`, and nonconstant candidate tasks `1 < 15`. The 64 anchor and 176 candidate records matched their expected cell counts, with zero infrastructure-missing cells and zero retries. This rules out absent cells and recorded infrastructure failures as the immediate cause of the block. It does not distinguish task difficulty from exact-system, scaffold, or protocol effects. Selection stopped before any Estimate or Diagnose computation; no evaluation or reference outcomes were accessed. Supplement S13 records these aggregate selection diagnostics, not per-cell trajectories.

The response matrix supplies no independent construct blueprint. Rank-preserving selection can erase semantic, safety, or accessibility coverage, favor incumbents, or invite gaming. The audit does not solve contamination, task or harness validity, licensing, or governance; its cost proxy excludes piloting, human review, provenance, drift monitoring, and abstention reruns. Evidence comes from coding-agent benchmarks. Workflow-level or causal claims require maintainer-blind future tasks, unseen scaffold families, and non-coding or cross-institutional tests.

12 REPRODUCIBILITY

Supplement S1–S12 contains the benchmark registry, objective-tradeoff and uncertainty tables, task and row robustness checks, frozen transfer results, claim-support material, and HFI-definition, cost, and chronology sensitivities. Supplement S13 contains the aggregate generation-6 tau2-airline selection lock; it includes neither per-cell trajectories nor evaluation/reference outcomes. The offline reproducibility workflow runs only local artifact-maintenance checks. A metric-source map and numeric-consistency audit link paper literals to locked paper-facing evidence tables. Anonymous supplementary materials also include processed matrices, code, tests, and derived evidence; redistribution follows upstream terms.

Ethics Statement

This retrospective analysis uses public benchmark result artifacts and does not involve human subjects. Results should nevertheless be interpreted carefully: leaderboard rows can share models, scaffolds, harnesses, training data, and submission practices. A maintenance policy can also change which capabilities receive attention. Outcome-derived diagnostics may nominate but must not override domain, safety, accessibility, licensing, or stakeholder review.

AI use statement

In this work, we used generative AI tools provided by OpenAI (Codex) and Anthropic (Claude Code) to assist with methodological feedback, experiment design, code implementation, execution and interpretation of offline analyses, literature discovery, manuscript drafting and editing, production of figures and research artifacts, and simulated peer review. The authors reviewed and verified the AI-assisted outputs, including code, analyses, numerical claims, citations, and manuscript text. Numerical claims were checked against the metric-source map and numeric-consistency audit described in the Reproducibility section. The authors take responsibility for the final content, claims, and artifacts.

References

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, Yiyang Nan, Jyoutir Raj, Yu Fan, Shubham Singh, Subramanyam Sahoo, Eliya Habba, Usman Gohar, Siddhesh Pawar, Robert Scholz, Arjun Subramonian, Jingwei Ni, Mykel Kochenderfer, Sanmi Koyejo, Mrinmaya Sachan, Stella Biderman, Zeerak Talat, Avijit Ghosh, and Irene Solaiman. When AI benchmarks plateau: A systematic study of benchmark saturation, 2026. URL https://arxiv.org/abs/2602.16763.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid language model benchmarking, 2025. URL https://arxiv.org/abs/2509.11106.

Wei-Jung Huang. How many tasks are enough for agent benchmark decisions? a replay analysis of public LLM agent benchmarks. KDD 2026 Workshop on Evaluation and Trustworthiness of Agentic AI, 2026. URL https://arxiv.org/abs/2607.12338.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for AI evaluation?, 2026. URL https://arxiv.org/abs/2607.15190.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues?, 2023. URL https://arxiv.org/abs/2310.06770.

Alex Kipnis, Konstantinos Voudouris, Luca Schulze Buschoff, and Eric Schulz. metabench: A sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4ebc26584810a189ef1e4f173aba4319-Abstract-Conference.html.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V. Chawla. Adaptive testing for LLM evaluation: A psychometric alternative to static benchmarks, 2025a. URL https://arxiv.org/abs/2511.04689.

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. PMLR, 2025b. URL https://proceedings.mlr.press/v267/li25bp.html.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents, 2023. URL https://arxiv.org/abs/2308.03688.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026. URL https://arxiv.org/abs/2601.11868.

Franck Ndzomga. Efficient benchmarking of AI agents, 2026. URL https://arxiv.org/abs/2603.23749.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation, 2026. URL https://arxiv.org/abs/2606.08679.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2519–2536. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.139. URL https://aclanthology.org/2024.naacl-long.139/.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices, 2024. URL https://arxiv.org/abs/2411.12990.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346/.

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, and Suparna Bhattacharya. SubLIME: Subset selection via rank correlation prediction for data-efficient LLM evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30572–30593. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1477. URL https://aclanthology.org/2025.acl-long.1477/.

Tommy Sha and Stella Zhao. BenchScope: How many independent signals does your benchmark provide?, 2026. URL https://arxiv.org/abs/2603.29357.

Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh Chawla. Benchmarkcards: Standardized documentation for large language model benchmarks, 2024. URL https://arxiv.org/abs/2410.12974.

SWE-bench. SWE-bench verified, 2024. URL https://www.swebench.com/verified.html.

Sang T. Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60238–60265. PMLR, 2025. URL https://proceedings.mlr.press/v267/truong25c.html.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1141–1158, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.92. URL https://aclanthology.org/2021.acl-long.92/.

Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, Xingzhang Ren, Fei Huang, Dayiheng Liu, and Linfeng Zhang. Rethinking LLM evaluation: Can we evaluate LLMs with 200x less data? In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.10457.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark, 2024. URL https://arxiv.org/abs/2406.19314.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Linghao Zhang, Shilin He, Chaoyun Zhang, Yu Kang, Bowen Li, Chengxing Xie, Junhao Wang, Maoquan Wang, Yufan Huang, Shengyu Fu, Elsie Nallipogu, Qingwei Lin, Yingnong Dang, Saravan Rajmohan, and Dongmei Zhang. SWE-bench goes live!, 2025. URL https://arxiv.org/abs/2505.23419.