SPARSE AGENT LEADERBOARDS: IDENTITY, SUPPORT, AND IDENTIFIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse agent leaderboards compare model–harness configurations under uneven coverage. A model ranking still needs one shared evaluation target. We map task identity, common-target support, and design identification onto three actions: certify a supported ordering, label a model-assisted extrapolation, or defer the comparison. Under task additivity, an incidence graph characterizes identifiable configuration contrasts, while free model–harness interactions leave model main effects unidentified. Sharp probability-scale bounds quantify what missing cells alone permit. In 22,818 archived Terminal-Bench trials spanning 53 configurations, candidate task-label joins connect a disconnected graph without adding any of the 4,702 observed cells, and a 55-record provenance sample contains four reported signatures within one name family. Under every label policy the best-covered model still lacks more than 60% of a uniform harness–task target, so no model pair can be certified against every bounded completion. On a separate public panel of 150 published aggregates, DeepSeek V3.2 beats GPT-5.2 under the source four-domain weights and loses under equal benchmark weights. At 90% observed cells the certificates cover 57.75–61.85% of pairwise orderings, while available-score ranking still publishes all ten pairs. Point estimates can match that coverage with almost no errors on this fixed table. The certificate is a worst-case column and a next-cell rule, not a rival accuracy score. Revealing ten additional deferred-weight cells on random 90% masks raises four-domain certificate coverage from 61.65% to 97.75%, versus 84.25% for ten random extra cells. A table-agnostic CSV tool returns the same three actions from observed cells.

1 INTRODUCTION

Agent performance depends on the model and the harness that supplies prompts, tools, control flow, retry rules, and budgets. Public and controlled studies demonstrate substantial harness effects (Zhang et al., 2026; Han & Sun, 2026; Yao et al., 2026). Evaluation infrastructure now makes these dependencies directly observable: HAL analyzes agents across models, scaffolds and benchmarks (Kapoor et al., 2025), while General Agent Evaluation crosses five models with five architectures and six benchmarks (Bandel et al., 2026). These studies establish the configuration as a meaningful evaluation unit.

A ranking for a fixed workflow needs more. Suppose model A is observed on harness H_1 with tasks T_1, T_2, while model B is observed on harness H_2 with T_2 only. Averaging available scores then compares different environments. A hierarchical fit can fill the missing cells (Gelman & Hill, 2006), but the filled ranking inherits the completion rule. The protocol asks three questions before that ranking is published. *Identity:* do shared task names name the same versioned task and scorer? *Support:* how much of the intended harness–task mixture was actually run for both models? *Identification:* if a statistical model completes the rest, which contrasts does the observed design recover? The reporting actions follow. If every bounded completion of the missing cells keeps the same sign, certify the ordering. If the sign depends on a fitted model, label the ranking as extrapolation. Otherwise defer and name the cell that would shrink the interval. Joining task aliases can connect a graph without adding observations; task-additive contrasts can then be identified while most target mass remains unseen. A sharp posterior resolves neither identity nor support. Figure 1 links these diagnostics to reporting. The distinction is already visible on a published panel (see Section 8). Under

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

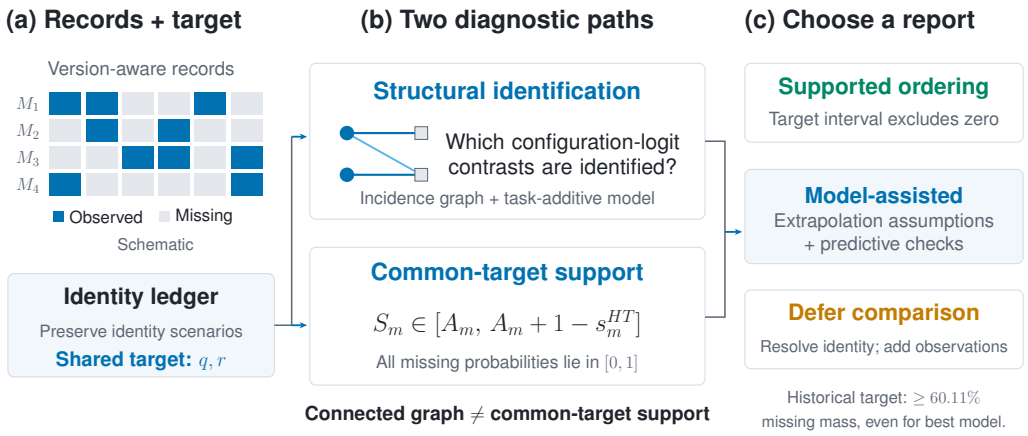

Figure 1: Two diagnostic paths for sparse comparisons. **(a)** Resolve identities and declare a shared target. **(b)** Task-additive logit identification and probability-scale support bounds. **(c)** Certify, label model-assisted extrapolation, or defer and acquire evidence. The matrix is schematic; 60.11% is the minimum missing mass on the historical candidate-joined uniform target.

the source four-domain weights DeepSeek V3.2 outscores GPT-5.2 (0.41026 versus 0.40554), while equal benchmark weights reverse the pair. On random 90% masks, available-score averaging still names a winner for every pair, while the certificate reports only 6.17 of ten.

Our contributions are:

(1) We propose a comparison procedure that maps identity status, declared target weights, graph identification and missing-support bounds to three reporting actions (certify, extrapolate, defer), with a table-agnostic CSV tool that returns those actions and a next cell (Sections 3–5).

(2) We conduct a reproducible case study in which candidate label joins restore graph connectivity without adding observations or resolving provenance (Section 6).

(3) We also conduct a public-panel study that applies the same certificate and next-cell rule to a separate crossed aggregate table and separates a worst-case ordering certificate from predictive point estimates (Section 8).

2 Related Work

Agent evaluation infrastructure and diagnosis. CurveShift (Xing et al., 2026) documents the retrospective attribution problem when model and harness deployment histories are confounded. Harness-Bench (Yao et al., 2026) instead constructs 106 controlled tasks and records 5,194 trajectories across model–harness configurations. It provides a controlled diagnostic benchmark. Our focus is the estimands and failure modes of statistical decomposition in an existing sparse leaderboard. The Scaffold Effect (Vats & Golev, 2026) complements this view with configuration-level efficiency and failure analysis on Terminal-Bench Pro.

Concurrent statistical decomposition. Han & Sun (2026) separate LLM and harness effects on Terminal-Bench 2.0 and report heterogeneity. Hardy et al. (2026) estimate generalizability for models and model–scaffold pairs on HAL, including decision-study projections. Those papers ask how much rank noise a random-facet design leaves. We ask which contrasts a fixed sparse panel identifies, which orderings survive every bounded completion of the missing cells, and which cell to acquire next. Reliability projections and likelihood identification are different targets.

Psychometrics and hierarchical measurement. Many-facet Rasch models (Linacre, 1989) and generalizability theory (Brennan, 2001; Jiang & Skorupski, 2018) are methodological ancestors. Empirical comparisons show that generalizability and many-facet Rasch analyses answer complementary questions in performance assessment, including sparse rating designs (Sudweeks et al., 2004; Smith & Kulikowich, 2004; Wind et al., 2023). Hierarchical response models extend this

perspective through partial pooling across persons, items, and facets (Fox, 2010, Chs. 2 and 6). The multilevel trifactor formulation provides a related Bayesian example (Fujimoto, 2019), while hierarchical regression supplies the broader modeling framework (Gelman & Hill, 2006). Our facets are engineering artefacts with version drift and evaluation costs. GLM identifiability under rank-deficient designs is classical (Fahrmeir & Tutz, 2001). Proposition 1 instantiates it for the edge–task incidence graph and separately identifies the role of the model–harness graph. Proposition 2 describes the remaining cross-component ambiguity, which priors or restrictions can resolve conditionally.

Partial identification and selective decisions. Missing-outcome bounds are classical (Manski, 2003), and inference for partially identified targets must distinguish the identification region from sampling uncertainty (Imbens & Manski, 2004). We apply this distinction to a declared model–harness–task mixture, rather than present the interval formula as a new estimator. Selective classification studies the trade-off between decision coverage and conditional error (El-Yaniv & Wiener, 2010). Our certificate is a set-valued comparison rule: it reports a sign only when every bounded completion agrees. Matched-coverage predictive baselines test empirical selective accuracy, which is distinct from this worst-case guarantee.

Harness optimisation. Meta-Harness (Lee et al., 2026) searches over scaffolds to *maximise* a model's score. Our framework concerns the targets and evidence used to compare configurations, including when allocation and reporting objectives differ. Wang et al. (2026) distinguish one-shot optimization gains from transfer and repeated optimization on newly introduced Terminal-Bench tasks. This motivates recording adaptation history and reserving tasks when evaluating evolving harnesses.

Benchmark integrity. Zhang et al. (2026) analyze harness disclosure through a position-paper formalisation and a small controlled factorial decomposition. Our case study tests whether task identity, support, and predictive adequacy justify interpretations of a larger sparse matrix. Harness taxonomies (Meng et al., 2026; Li et al., 2026), reproducible environments (Garg et al., 2025) and configuration-level safety evaluation (Bai et al., 2026) further motivate recording execution and adaptation history alongside scores.

3 Comparison Targets and Measurement Assumptions

Let y_{mht} be the number of successes in $n_{mht} > 0$ recorded trials for model m, harness h, and task t. An entity denotes a versioned evaluation object when its identity is known, otherwise we explicitly use a reported label. The observed cell set is $\mathcal{O}$. A baseline working likelihood is $y_{mht} \sim \mathrm{Binomial}(n_{mht}, p_{mht})$, with

$$\eta_{mht} = \mathrm{logit}(p_{mht}) = \mu + \alpha_m + \beta_h + \gamma_{mh} - \delta_t = \theta_{mh} - \delta_t. \tag{1}$$

The task-additive restriction means that a task has the same logit-scale difficulty for every configuration. It is a substantive model assumption.

A shared probability target. Fix common distributions q over harnesses and r over tasks, including their versions, before comparing models. The expected success rate is

$$S_m(q, r) = \sum_{h,t} q_h r_t p_{mht}, \qquad \Delta_{ij}(q, r) = S_i(q, r) - S_j(q, r). \tag{2}$$

A non-product deployment distribution can replace $q_h r_t$ throughout. This target depends on the declared environment. Averaging probabilities makes its interpretation independent of how a predictor is partitioned into effects. An unobserved positive-weight cell still requires extrapolation or bounds.

Choosing an actionable target. A deployment comparison can fix one shared harness, a planned mixture of compatible harnesses, or a benchmark-wide reference mixture. These answer different questions. We use uniform weights as a coverage stress test, whereas deployment weights should encode the intended environment. Compatibility is fixed before scores are inspected. An incompatible configuration either lies outside a prespecified common feasible target or receives the failure outcome required by the deployment policy. Dropping it separately for each model silently changes the comparison. Report the target, feasibility rule and a small set of prespecified weight sensitivities together.

Support of that target. The quantity that enters the reporting rule is the joint support

$$s_m^{HT}(q, r) = \sum_{(h,t):(m,h,t)\in\mathcal{O}} q_h r_t, \tag{3}$$

the target mass sitting on observed cells. If O_m^H is the observed harness set, harness-only support $s_m^H(q) = \sum_{h\in O_m^H} q_h$ can be larger, because an observed configuration need not cover every target task. Archived tables keep $s_m(q) = s_m^H(q)$. Renormalizing weights separately for each model changes the target. For the archived logit fit we also use

$$C_m(q) = \sum_h q_h \theta_{mh} \tag{4}$$

Because logit^{-1} is nonlinear, C and S need not rank models in the same order.

Historical inference. The archived hierarchy uses independent zero-mean Gaussian effects with HalfNormal(1) scale priors, $\mu \sim \mathcal{N}(0, 2)$, and non-centred NUTS inference (Hoffman & Gelman, 2014). Its main fit uses four chains with 4,000 warmup and 4,000 retained draws each. These priors regularize a non-unique decomposition. Their hypervariances provide conditional logit-scale attributions under this parameterization.

4 Identification in Sparse Designs

Let $\mathcal{E}_O$ contain observed model–harness pairs. The bipartite graph $\mathcal{B}$ joins those configurations to observed tasks, while $\mathcal{G}$ joins models to harnesses. Claims below condition on the chosen identity map.

In words, under task additivity: (i) connected observed cells identify contrasts among those configurations; (ii) a comparison that crosses disconnected components needs an extra restriction; (iii) if every model–harness interaction is free, model main effects are not recovered from the likelihood. Proposition 1 is the corresponding rank statement, and the proof is in Appendix C.

Proposition 1 (Task-additive identification)**.** *For an injective link and positive observed trial counts, the sampling distribution under Equation 1 identifies within-component differences of configuration effects and of task effects, up to one common location per component of $\mathcal{B}$. A linear contrast $L = \sum_{e\in\mathcal{E}_O} w_e\theta_e$ is identified exactly when $\sum_{e\in C} w_e = 0$ in every component C. A nonzero coefficient on an unobserved configuration is unsupported under the free-edge model. With free γ_{mh}, $\alpha_i - \alpha_j$ is not conditionally likelihood-identified. If $\gamma = 0$, connected $\mathcal{B}$ and $\mathcal{G}$ are sufficient for additive model and harness contrasts after location constraints.*

Proposition 2 (Disconnected comparisons)**.** *Under the free-edge/task-additive model, independent joint shifts of configuration and task effects in different components preserve the likelihood. Cross-component location comparisons consequently require additional restrictions or priors.*

For a $C_i(q) - C_j(q)$ contrast, both models need observed configurations on the support of q, with equal component-wise q mass (Appendix K).

Model–task interactions change the identification problem. Consider $\eta_{mht} = \theta_{mh} - \delta_t + \zeta_{mt}$. Conditional on unrestricted realized task interactions, the transformation

$$\theta'_{mh} = \theta_{mh} + b_m, \qquad \zeta'_{mt} = \zeta_{mt} - b_m \tag{5}$$

preserves every probability but shifts $C_m(q)$ by b_m. Thus the graph certificate for configuration logits depends on task additivity. A zero-mean random-effect distribution or a weighted constraint on ζ adds structure. Identification after integrating random effects is a different question from conditional likelihood identification. Equation 2 is invariant to Equation 5, but still requires its own support. Our executable four-model example preserves predictors to numerical precision.

5 A Support-Aware Comparison Procedure

The procedure records distinct evidence states and maps them to reporting actions.

(1) **Resolve identities.** Preserve raw names, task content and scorer identifiers, source revisions, and model/harness configurations. Evaluate an unresolved alias through explicit sensitivity scenarios.

(2) **Declare the target.** Fix the common q, r and the comparison scale. Compute both harness and joint support.

(3) **Check structure.** Under task additivity, test the null-space condition for each requested logit contrast. Record unsupported edges and component-location dependence separately.

(4) **Assess estimation.** For a chosen model, report sampling diagnostics, predictive checks and validation matched to the intended prediction unit. Compare extrapolations across justified specifications.

(5) **Choose an action.** Certify a supported sign, label a model-assisted prediction, or defer and name the next cell.

The implementation releases candidate identity maps, per-policy graphs, pairwise structural certificates and support diagnostics. Structural certificates state their statistical model, while calibration is assessed by separate predictive checks.

An identity ledger has three outcomes: *verified equivalent* when task content, scorer and relevant execution settings agree; *substantively different* when an outcome-relevant change is established; and *unresolved* otherwise. A discrepant task signature triggers content and scorer inspection. Candidate maps for unresolved records remain separate analysis scenarios.

Probability-scale bounds without extrapolating missing cells. Let $A_m = \sum_{(h,t):(m,h,t)\in\mathcal{O}} q_h r_t p_{mht}$. If only $0 \le p_{mht} \le 1$ is imposed on missing cells, then

$$S_m \in [A_m,\ A_m + 1 - s_m^{HT}], \quad \Delta_{ij} \in [A_i - A_j - (1 - s_j^{HT}),\ A_i - A_j + (1 - s_i^{HT})]. \tag{6}$$

These are sharp missing-outcome bounds in the unrestricted probability model (Manski, 2003): assigning all missing probabilities to zero or one reaches the endpoints, while intermediate assignments fill the intervals. Their width is a support property, even when the probabilities on observed cells are known exactly. Structural restrictions can narrow the set, but the resulting bounds then inherit those restrictions. The bound turns observed support directly into an interpretable reporting quantity.

Finite-sample uncertainty in A_m is separate. The released plug-in endpoints condition on observed proportions and are not confidence intervals. Under justified independent binomial sampling, simultaneous cell intervals can be propagated through the positive weights and combined with the missing-support term. The retrospective predictive diagnostics below motivate calibration checks beyond the working binomial likelihood.

For $O_m = \{(h,t) : (m,h,t) \in \mathcal{O}\}$, let $[\ell_{mht}, u_{mht}]$ jointly cover observed-cell probabilities with probability at least $1 - \varepsilon$. Define

$$L_m = \sum_{(h,t)\in O_m} q_h r_t \ell_{mht}, \qquad U_m = \sum_{(h,t)\in O_m} q_h r_t u_{mht} + 1 - s_m^{HT}. \tag{7}$$

Then $[L_i - U_j, U_i - L_j]$ covers every requested pairwise contrast on the same simultaneous event. Allocating a total error budget ε across valid cell intervals is one conservative construction. Dependence or overdispersion requires intervals valid under the corresponding sampling process (Imbens & Manski, 2004). This separates the reducible sampling term from missing-support uncertainty and from structural identification.

From an interval to an evaluation action. Certify $i > j$ only when the lower endpoint for Δ_{ij} is positive, while certify $j > i$ only when the upper endpoint is negative. Otherwise defer an unrestricted ordering or attach explicit assumptions to a model-assisted one. If observed A_m are treated as known, the pair interval has width $(1 - s_i^{HT}) + (1 - s_j^{HT})$. A newly revealed cell of weight $q_h r_t$ for model i cuts the first term by that weight. The sign is certified only after the interval no longer contains 0. Repeating an already observed cell changes sampling precision but not support mass, so graph-bridging evaluations, high-target-mass missing cells, and independent repeats repair different gaps.

Table 1: Identity-policy sensitivity on the same archived cells. Candidate joins define scenarios. Entries are deterministic for the frozen snapshot and each policy's label grid.

Task-label policy	Labels	$\mathcal{B}$ components	Grid density
Literal archived labels	179	2	5.15%
Remove namespace prefix (candidate)	90	1	10.24%
Also join Windows names (candidate)	89	1	10.36%

6 Retrospective Evidence: Identities and Support

Population and identity policy. The assembled inventory has 38,333 records: 32,803 Terminal-Bench trials, 323 SWE-bench aggregates, and 5,207 METR rows. The historical fit selects a Terminal-Bench model–harness component with 17 model labels, 30 harness labels, 53 configurations, 4,702 cells, and 22,818 trials. Source selection and outcome scales are described in Appendix N. This is a retrospective feasibility sample.

The historical task variable contains 179 literal labels. Its two graph components use 89 bare names and 90 prefixed names. Removing `terminal-bench/` matches all 89 bare names. The remaining distinction is `install-windows-3-11` versus `install-windows-3.11`. candidate label joins improve conditional graph connectivity, while identity verification and target coverage require separate evidence (Figure 8, Appendix O).

Table 1 reports the structural consequences of these candidate joins. All policies preserve 22,818 trials, 15,372 successes, and 4,702 cells. Each structural interpretation therefore states its identity policy.

Conditional on task additivity, the literal-label graph identifies 1,138 of the $\binom{53}{2} = 1,378$ observed-configuration contrasts. Either candidate join makes all 1,378 structurally identifiable. This contrasts with the unsupported full-harness model comparisons below: observed-configuration contrasts and extrapolated model rankings are different targets.

Public records contain multiple task signatures. We fixed the public dataset snapshot accessed on 7 September 2026 and downloaded one lexicographically selected trial from each of the 55 fitted submissions. All 55 samples belong to the adaptive-rejection-sampler name family: 49 bare and six prefixed labels. They report four distinct task signatures, with counts 48, 5, 1, and 1. One prefixed record declares `terminal-bench/terminal-bench-2-1`, while others use a 2.0 source identifier or a local dataset path. Signature differences can reflect packaging conventions as well as substantive task changes. The sample therefore establishes multiple serialized representations within one name family. It is not a corpus-wide identity audit. Semantic equivalence requires content- and scorer-level reconciliation. Redacted evidence summaries, selection rules, source URLs and retrieval metadata are released.

The probe represents the fixed public release and motivates two complementary views: the literal-label fit preserves the historical analysis, while joined graphs measure structural sensitivity to candidate identity maps. A version-resolved posterior requires content- and scorer-level reconciliation across the full corpus.

Support remains sparse under every candidate map. Each model appears with 1–12 of 30 harness labels, giving uniform harness support of 3.3–40.0%. No harness is observed for every model. For uniform weights over each policy's harness–task grid, joint support ranges from 1.62–19.83% under literal labels, 3.22–39.44% after the prefix join, and 3.26–39.89% after both candidate joins (Table 2). These are policy-specific target populations, not interchangeable estimates of one deployment distribution.

In every policy, every model has $s_m^{HT} < 0.4$. Thus every model's support interval in Equation 6 contains $1/2$, regardless of the observed-cell probabilities. No one of the 136 model-pair orderings can be certified uniformly over unrestricted completions of the missing cells. This certificate concerns robustness to all bounded completions, and structured models and targeted observations provide additional routes to informative decisions.

Table 2: Joint support for uniform harness–task targets. The last column is the smallest missing mass among 17 models. Entries are deterministic for the frozen observation matrix and declared weights. Sampling uncertainty is separate.

Task-label policy	Minimum support	Maximum support	Min. unresolved
Literal	1.62%	19.83%	80.17%
Prefix join (candidate)	3.22%	39.44%	60.56%
Prefix + Windows (candidate)	3.26%	39.89%	60.11%

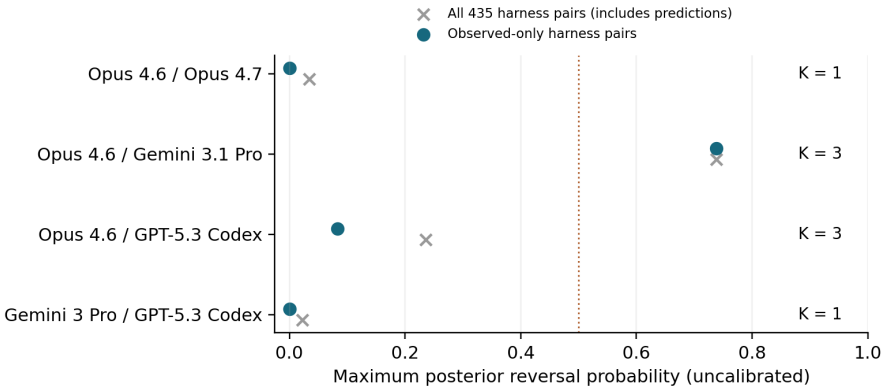

Figure 2: Observed-supported crossover screen. Teal: jointly observed harnesses only. Gray: all 435 harness pairs, including predictions. The 0.50 mark is descriptive.

Figure 2 is the observed-supported screen. Teal points restrict each contrast to harnesses run for both models. The certificate instead uses every bounded completion of the missing target mass in Equation 6. Gray crosses that score all 435 harness pairs are model-assisted extrapolations. The numerical observed-only maximum is in Table 5. Its working-model scope is in Appendix K.

7 Historical Model Adequacy and Predictive Evidence

The following diagnostics use the literal-label population. Candidate-join refits would define different task parameters, residuals and task-block uncertainty.

Aggregate fit conceals cell-level dispersion. The baseline reproduces aggregate model, harness and task success rates inside marginal 90% predictive intervals, and its marginal cell intervals cover 92.15% of observed counts. Yet Table 3 shows substantial unreproduced dispersion. Adding $\zeta_{mt} \sim \mathcal{N}(0, \sigma^2_{mt})$ estimates $\sigma_{mt} = 1.598$ (Table 16) and reduces the discrepancy to 7,715 versus 4,561. Residual overdispersion remains. The observed-edge-only result places the same discrepancy inside the observed design.

Direct predictive comparison has a narrower scope. Nine completed refits compare the full, no-interaction and no-harness structures on a shared three-fold held-out-cell partition. All have maximum rank-normalized $\widehat{R} < 1.01$ (Vehtari et al., 2021) and zero divergences. Relative to full, removing interaction changes total held-out elpd by -0.46 (cell SE 6.23, literal-task-block SE 6.25), and removing the harness main effect changes it by -6.51 (SEs 3.46 and 3.62). These results provide little predictive discrimination among the original structures. Their evidential scope is prediction for omitted cells within the observed design. The model–task extension is outside this refit comparison. Its PSIS diagnostics flag 516/4,702 cells with $k > 0.7$.

Identity and fit uncertainty remain distinct. The strict metadata subset has 4,998 trials over eight model labels, five harness labels, and 89 task labels. Its eight-model index ordering agrees with the corresponding baseline ordering, while residual discrepancy remains. The more granular reported-

Table 3: Predictive discrepancy in historical literal-label fits. Entries are median observed/replicated Pearson discrepancies. Upper-tail frequencies and exact 95% Monte Carlo confidence intervals (MC CIs) use 2,000 draw-matched replicates; the CIs quantify simulation precision.

Specification	Observed	Replicated	Upper tail	95% MC CI
Baseline	12,872	4,679	0	$[0, 0.0018]$
Model–task interaction	7,715	4,561	0.0075	$[0.0042, 0.0123]$
Observed configurations only	12,962	4,682	0	$[0, 0.0018]$
Strict metadata subset	2,102	1,128	0.0050	$[0.0024, 0.0092]$

Table 4: Public-panel masking, four-domain target: mean $\pm$ sample SD (percentage points) across 200 masks. Available and ridge report all pairs in every mask shown, so mean seed-level error equals pooled error. Certificate errors against the fixed table are zero by bounded completion.

Observed	Mask	Cert. coverage	Available error	Ridge error
80%	Random	25.50 ± 10.50	5.45 ± 7.00	4.20 ± 5.05
	Config. block	28.05 ± 10.50	5.05 ± 6.01	5.25 ± 6.01
	Score-dependent	10.05 ± 8.30	7.10 ± 7.20	5.70 ± 6.69
90%	Random	61.65 ± 6.86	3.55 ± 5.39	3.50 ± 4.78
	Config. block	61.85 ± 9.67	3.45 ± 4.97	3.00 ± 4.59
	Score-dependent	57.75 ± 7.66	7.80 ± 7.58	5.30 ± 6.57

identity partition has maximum $\widehat{R} = 1.0136$. Appendix L presents it as a convergence-qualified sensitivity.

The historical ratio $(\sigma_\beta^2 + \sigma_\gamma^2)/\sigma_\alpha^2$ has median 0.40 (90% CrI in Table 7). Appendix A states its literal identity policy, hierarchy and priors.

8 Public Crossed-Panel Test: When Does Abstention Inform?

Data and reference target. We reanalyze the public Open Agent Leaderboard associated with Bandel et al. (2026), using the snapshot accessed on 7 September 2026. It supplies all 150 aggregate scores for five models, five agent architectures and six benchmarks. Here architectures instantiate the harness axis and benchmarks replace individual tasks. We preserve the source's primary-score convention and retain every row, including configurations with incomplete sessions. The complete finite table defines the reference weighted score. We weight architectures equally and follow the source's four-domain weighting: AppWorld, BrowseComp+ and SWE-bench each receive $1/4$, and the three τ^2 domains each receive $1/12$.

Frozen masking comparison. For each of 200 seeds we reveal 20–100% of the table under random-cell, configuration-block and score-dependent masks. Acquisition conclusions use the first two; the third mimics publication bias. Available-score averaging renormalizes observed weights. Additive ridge imputes missing scores from model, architecture and benchmark effects. Coverage is the fraction of ten pairwise orderings reported; error is disagreement with the full-table sign. SD summarizes mask variation on this fixed panel. Matched-coverage baselines report exactly as many pairs as the certificate in each mask, ordered by predicted margin.

Useful certificates require sufficient support and separation. At 60% observed cells, certificate coverage is only 0.05, 2.50 and 0% for random, block and score-dependent masks. At 90%, it rises to 57.75–61.85% (Table 4), compared with zero for always abstaining. A simple support gate is different: at 90% random observation, ridge with minimum pair support 0.9 reports 30.45% of pairs with 2.96% error. Certificates use the observed margin as well as support. At exactly the certificate's per-mask decision count, both margin-ranked baselines make zero errors for the four-domain target at 80% and 90% observation under all three masks (Appendix P). On the six-benchmark uniform target at 90% observation, the same matched-coverage rule yields 1–2 baseline errors under block and score-dependent masks and zero certificate errors. That is reliability of a point estimate on this table, not a reason to collapse the two columns: available-score ranking still publishes all ten pairs

Table 5: The same scores authorize different actions once the target and evidence state are declared. Source marks the corpus. Public-panel values are exact finite-table contrasts, not future-run significance.

Source	Setting	Supported evidence	Reporting action
Public panel	DeepSeek V3.2 vs GPT-5.2, four-domain ($0.41026 > 0.40554$)	Full-table sign $+$; gap 0.005	Report the declared-target contrast
Public panel	Same pair, six-benchmark uniform ($0.46239 < 0.46636$)	Opposite full-table sign	Changed target, not a missing-data method
Retrospective archive	Opus 4.6 vs Gemini 3.1 Pro Preview	One observed-only maximum 0.739; only four pairs share ≥ 2 harnesses	Report the observed crossover; label remaining ranks as extrapolation

Table 6: Four-domain public panel, frozen random masks (means over 200 seeds). Policies choose ten cells without reading hidden scores.

Start observed	Start cov.	$+10$ deferred-wt.	$+10$ largest-wt.	$+10$ random
90%	61.65%	97.75%	92.15%	84.25%
80%	25.50%	59.95%	59.95%	51.65%

at 90% random observation, while the certificate reports 6.17. The withheld pairs are exactly the ones whose sign is not invariant to bounded completion. Uniform six-benchmark weighting gives 62.10–65.65% certificate coverage at 90% observation and reverses the DeepSeek V3.2 versus GPT-5.2 pair (Table 5). The reversal is a changed target, not a missing-data method. **Which missing cell is worth acquiring.** Certificates also rank the next evaluation. From each existing 80% and 90% mask we reveal 10 further cells under policies that do not inspect hidden scores: random remaining cells, largest target weight, or largest weight among models that still participate in a deferred pair. Revealed values are the published aggregates. On four-domain random 90% masks, available-score ranking reports all 10 pairs, the certificate reports 6.17, and ten deferred-weight cells raise the certificate to 9.78 pairs (97.75%), versus 84.25% for ten random extra cells (Table 6). The same $+10$ rule on 80% random masks raises coverage from 25.50% to 59.95%. The increments describe missingness design on this fixed table. Those three counts are the leaderboard columns: fill every pair, withhold uncertified pairs, and pick the next cell among models that still have a deferred pair, by target mass $q_h r_t$ and without reading the hidden score.

9 Discussion

The framework separates unresolved provenance, insufficient common-target support, selective ordering certificates, and predictive discrepancies. Resolve identity before pooling, acquire missing target mass before demanding unrestricted rankings, and label model-assisted estimates with their prediction unit. A release can carry versioned model, harness, task, scorer, budget and adaptation fields, a coverage matrix with target weights, and three columns: a supported sign, a point estimate, and the deferred pairs whose residual interval width ranks the next cell. The supplement's `sparse_lb` tool writes those columns from a ragged CSV of observed cells. It does not fit a hierarchy or require a complete grid. A sign change across prespecified targets is a changed decision problem, not one universal order. The next-cell rule follows residual support-bound width on the public panel. Appendix B is a semi-synthetic allocation stress test.

10 Conclusion

Sparse agent comparisons require identity resolution, common-target support and validated modeling assumptions. Graph identification characterizes one stated model. Probability support bounds reveal what remains unresolved without extrapolation. The public-panel study measures how much ordering those bounds retain as support grows, and which missing cells convert deferred pairs into certificates. Reporting identity status, target weights and decision coverage separates observed evidence from extrapolation, and tells the next evaluation whether to bridge a component, add target mass, or reduce sampling uncertainty.

Ethics Statement

This study analyzes public benchmark records and simulated outcomes. Its principal risks are incorrect entity attribution and overstated comparisons. We preserve raw labels and provenance, distinguish candidate joins from verified identities, and report model-dependent estimates with their scope.

AI Use Statement

Generative AI tools assisted with coding, editing, and reviewing. The authors conceived the original idea, built the research plan and theoretical framework, and performed the experiments and analysis. The authors reviewed all AI-assisted work and take responsibility for the final manuscript, claims, code, and artifacts.

Reproducibility and Data Availability

The anonymous supplement contains the archived data and analysis artifacts, task-policy and support code. It also includes `sparse_lb`, a CSV protocol that emits support, pairwise certificates and a deferred-weight next cell on rectangular or ragged tables, reproducing the public-panel certificate rule without a fitted hierarchy.

References

C. Agrell and K. Rognlien Dahl. Sequential Bayesian optimal experimental design for structural reliability analysis. *Statistics and Computing*, 31(3):27, 2021. doi:10.1007/s11222-021-10000-2.

Y. Bai, J. Duan, J. Peng, X. Wu, S. Liu, S. Wang, and T. Chen. HarnessRisk: A lifecycle-oriented benchmark for agent harness safety. arXiv preprint arXiv:2608.17597, 2026. doi:10.48550/arXiv.2608.17597.

E. Bandel, A. Yehudai, L. Eden, Y. Sagron, Y. Perlitz, E. Venezian, N. Razinkov, N. Ergas, S. S. Ifergan, S. Shlomov, M. Jacovi, L. Choshen, L. Ein-Dor, Y. Katz, and M. Shmueli-Scheuer. General agent evaluation. arXiv preprint arXiv:2602.22953, version 2, 2026. doi:10.48550/arXiv.2602.22953.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi:10.1111/j.2517-6161.1995.tb02031.x.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi:10.1080/01621459.2017.1285773.

R. L. Brennan. *Generalizability Theory*. Springer, 2001. doi:10.1007/978-1-4757-3456-0.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi:10.1214/ss/1177009939.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. `https://www.jmlr.org/papers/v11/el-yaniv10a.html`.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 2nd edition, 2001. doi:10.1007/978-1-4757-3454-6.

J.-P. Fox. *Bayesian Item Response Modeling: Theory and Applications*. Springer, 2010. doi:10.1007/978-1-4419-0742-4. See Ch. 2, pp. 31–44, and Ch. 6, pp. 141–191.

K. Fujimoto. The Bayesian multilevel trifactor item response theory model. *Educational and Psychological Measurement*, 79(3):462–494, 2019. doi:10.1177/0013164418806694.

D. Garg, D. Caples, A. Draguns, N. Ravi, et al. REAL: Benchmarking autonomous agents on deterministic simulations of real websites. *Advances in Neural Information Processing Systems*, 38, Datasets and Benchmarks Track, 2025. `https://arxiv.org/abs/2504.11543`.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006. doi:10.1017/CBO9780511790942.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24:997–1016, 2014. doi:10.1007/s11222-013-9416-2.

J. Han and Y. Sun. How good is your harness? Statistical evaluation of coding harnesses. In *1st Workshop on Combining Theory and Benchmarks (CTB@ICML 2026)*, 2026. https://openreview.net/forum?id=QI8z3skBwt.

G. W. Imbens and C. F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004. doi:10.1111/j.1468-0262.2004.00555.x.

M. Hardy, A. Reuel, R. Azam, M. Kochenderfer, and S. Koyejo. How reliable are agent leaderboards? A variance-decomposition analysis. RLEval workshop at ACM CAIS, 2026. https://openreview.net/forum?id=GAu4q6OkLI.

M. D. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014. https://jmlr.org/papers/v15/hoffman14a.html.

Z. Jiang and W. Skorupski. A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods*, 50(6):2193–2214, 2018. doi:10.3758/s13428-017-0986-3.

S. Kapoor et al. Holistic Agent Leaderboard: The missing infrastructure for AI agent evaluation. arXiv preprint arXiv:2510.11977, 2025. doi:10.48550/arXiv.2510.11977.

S. Kleinegesse, C. Drovandi, and M. U. Gutmann. Sequential Bayesian experimental design for implicit models via mutual information. *Bayesian Analysis*, 16(3):773–802, 2021. doi:10.1214/20-BA1225.

Y. Lee, R. Nair, Q. Zhang, K. Lee, O. Khattab, and C. Finn. Meta-Harness: End-to-end optimization of model harnesses. arXiv preprint arXiv:2603.28052, 2026. doi:10.48550/arXiv.2603.28052.

J. Li, X. Xiao, Y. Zhang, C. Liu, et al. Agent harness engineering: A survey. Withdrawn TMLR submission on OpenReview, 2026. https://openreview.net/forum?id=eONq7FdiHa.

J. M. Linacre. *Many-facet Rasch Measurement*. MESA Press, 1989. ISBN 0-941938-02-6. https://rasch.org/facet.htm.

C. F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003. doi:10.1007/b97478.

Q. Meng, Y. Wang, L. Chen, Y. Li, W. Wu, W. Jiang, Q. Wang, C. Lu, Y. Gao, Y. Wu, and Y. Hu. Agent harness for large language model agents: A survey. Preprints.org, version 3, 2026. doi:10.20944/preprints202604.0428.v3.

M. A. Merrill, A. G. Shaw, N. Carlini, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. arXiv preprint arXiv:2601.11868, 2026. doi:10.48550/arXiv.2601.11868.

METR. Time Horizon 1.1. Public release, January 29, 2026. https://metr.org/blog/2026-1-29-time-horizon-1-1/; public data repository: https://github.com/METR/eval-analysis-public.

S. Oladyshkin and W. Nowak. The connection between Bayesian inference and information theory for model selection, information gain and experimental design. *Entropy*, 21(11):1081, 2019. doi:10.3390/e21111081.

E. Pauwels, C. Lajaunie, and J.-P. Vert. A Bayesian active learning strategy for sequential experimental design in systems biology. *BMC Systems Biology*, 8:102, 2014. doi:10.1186/s12918-014-0102-6.

E. V. Smith and J. M. Kulikowich. An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, 64(4):617–639, 2004. doi:10.1177/0013164404263876.

R. R. Sudweeks, S. Reeve, and W. S. Bradshaw. A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3):239–261, 2004. doi:10.1016/j.asw.2004.11.001.

N. Vats and O. Golev. The scaffold effect in coding agents: Harness choice as a hidden variable in coding-agent evaluation. arXiv preprint arXiv:2607.22585, 2026. doi:10.48550/arXiv.2607.22585.

A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. doi:10.1007/s11222-016-9696-4.

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021. doi:10.1214/20-BA1221.

W. Wang and N. M. Kingston. Adaptive testing with a hierarchical item response theory model. *Applied Psychological Measurement*, 43(1):51–67, 2019. doi:10.1177/0146621618765714.

W. Wang, P. Kattakinda, and S. Feizi. Do agent optimizers compound? A continual-learning evaluation on Terminal-Bench 2.0. arXiv preprint arXiv:2607.14004, 2026. doi:10.48550/arXiv.2607.14004.

S. A. Wind, J. Jones, and S. Grajeda. Does sparseness matter? Examining the use of generalizability theory and many-facet Rasch measurement in sparse rating designs. *Applied Psychological Measurement*, 47(5–6):351–364, 2023. doi:10.1177/01466216231182148.

H. Xing, P. Wang, B. Meng, K. Alhamoud, X. Li, J. Wang, X. Yu, X. Han, X. Li, P. Torr, and Y. Hao. CurveShift: Is agent progress scalar? Separating level from shape. arXiv preprint arXiv:2608.00355, 2026. doi:10.48550/arXiv.2608.00355.

Z. Xu and Q. Liao. Gaussian process based expected information gain computation for Bayesian optimal design. *Entropy*, 22(2):258, 2020. doi:10.3390/e22020258.

Y. Yao, X. Tan, C.-H. Liu, Y. Li, Z. Wang, W. Yu, Z. Tan, Y. Tian, G. Zhao, L. Sun, X. Zhang, and T. Yang. Harness-Bench: Measuring harness effects across models in realistic agent workflows. arXiv preprint arXiv:2605.27922, 2026. doi:10.48550/arXiv.2605.27922.

Y. Zhang, J. Wang, Y. Ge, W. Xu, J. Hamm, and C. K. Reddy. Stop comparing LLM agents without disclosing the harness. arXiv preprint arXiv:2605.23950, 2026. doi:10.48550/arXiv.2605.23950.

A Archived Literal-Label Fit: Conditional Summaries

This appendix preserves measured historical results. Every task count and component statement below uses literal task labels, rather than validated task-version identities. Section 6 supersedes the earlier interpretation of two genuinely disjoint task sets. No candidate-join refit is represented by these numbers.

Data. Our corpus inventory contains 38,333 rows: 32,803 Terminal-Bench 2.0 trial records from 75 submissions (Merrill et al., 2026), 323 aggregate SWE-bench leaderboard rows, and 5,207 METR Time Horizon 1.1 task-level rows (METR, 2026). The statistical fit uses only the dominant Terminal-Bench component: 17 models, 30 harnesses, 179 task labels, 4,702 observed cells, and 22,818 trials. This is 5.15% of the nominal $17 \times 30 \times 179$ grid. The 38,333-row inventory is therefore not the effective sample size of the fitted model. Although the selected model–harness graph $\mathcal{G}$ is connected, its 53 observed configuration edges split into two literal-label components in $\mathcal{B}$: one contains 48 edges, 89 task labels, 4,266 cells, and 20,379 trials; the other contains 5 edges, 90 task labels, 436 cells, and 2,439 trials. Their relative locations are prior-dependent under this particular identity policy and free-interaction model. Candidate namespace joins connect this graph; their version validity remains unresolved (Section 6).

Identity provenance. Public records do not uniformly expose exact model and harness versions. Among 27,948 otherwise eligible single-model Terminal-Bench trials, 8,345 (29.9%) lack a raw model identifier, 20,415 (73.0%) have empty or brace-only harness-detail metadata, and 1,326 use one of three curated model aliases. These categories overlap and flag incomplete metadata, not known errors. A conservative complete-metadata filter leaves a different 8-model, 5-harness, 89-task component with 4,998 trials, so it cannot serve as an apples-to-apples robustness fit. We release the record-level identity map and treat version alignment as a limitation of the retrospective evidence. Appendix L tests a metadata-complete subset and a more granular identity partition as changed analysis populations, not recovered true versions.

Fit. NUTS 4 chains $\times$ (4000 warmup + 4000 sampling) on CPU; $\max \widehat{R} = 1.0074$ (gate < 1.01), min ESS $= 1{,}043$, 0 divergences, 12.5 minutes.

Archived crossover statistic. For model pair (i, j), the screen is

$$p_{ij} = \max_{h,h'} \Pr[(\theta_{ih} - \theta_{jh})(\theta_{ih'} - \theta_{jh'}) < 0 \mid \mathcal{D}] \,.$$

Table 7: Archived literal-label conditional variance attribution on the Terminal-Bench dominant component (posterior means of σ^2, posterior median for the ratio, and 90% posterior credible intervals, CrIs). Under the stated prior-regularised parameterisation, the posterior harness variance is about 35% of the posterior model variance. The interaction is small in this fit, but the aliasing analysis (Table 8) shows that it absorbs harness variance when β is removed. These are parameterisation-dependent attributions, not likelihood-identified effects.

Component	Posterior estimate	90% CrI	Share of σ^2_α
Model σ^2_α	1.348	[0.69, 2.35]	—
Harness σ^2_β	0.470	[0.24, 0.78]	35%
Interaction σ^2_γ	0.057	[0.007, 0.154]	4%
Task σ^2_δ	4.913	[3.99, 6.00]	—
$(\sigma^2_\beta + \sigma^2_\gamma)/\sigma^2_\alpha$	0.40 (median)	[0.18, 0.92]	$\Pr(\geq 0.3) = 0.72$

The observed-only version restricts the maximum to four observed configurations with connected within-harness contrasts under the literal-label task-additive model. The 0.50 threshold is descriptive. The quantity $1 - p_{ij}$ is not a frequentist p-value, so a Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) does not confer FDR control on this screen.

Reading. The analysis-plan ratio $(\sigma^2_\beta + \sigma^2_\gamma)/\sigma^2_\alpha$ has posterior median 0.40, with its 90% posterior interval in Table 7, and probability 0.72 of exceeding 0.3. This is an attribution within a working model. Its adequacy must be checked before substantive interpretation. Across 2,000 posterior predictive replicates, all model, harness, task, component, and trial-count aggregate rates fall within their marginal 90% predictive intervals. Nevertheless, median observed Pearson discrepancy is 12,872 against 4,679 in replicated data; none of the replicates exceeds its draw-matched observed discrepancy. Aggregate calibration therefore conceals substantial cell-level misfit. Adding a pooled model–task term gives $\sigma_{mt} = 1.60$ (Table 16) and changes the model-index order to $\tau = 0.941$ relative to baseline; residual dispersion remains (Appendix L).

Support and crossovers. Each model is observed with 1–12 of the 30 harnesses: uniform q assigns 60.0–96.7% of its mass to missing edges. There is no common observed harness across all 17 models. Restriction to each model's own support changes the posterior-mean order ($\tau = 0.897$, maximum shift 2), while changing the estimand. Only 4 of 136 model pairs have two jointly observed harnesses and hence an observed-only crossover comparison; one exceeds 0.50. The archived extrapolative screen has two exceedances. These different denominators and supports preclude a "2/136 observed crossover rate" reading (Appendix K).

Ablation and robustness (Table 8). Removing the interaction ($\gamma \equiv 0$) or the harness main effect ($\beta \equiv 0$) changes leave-one-out predictive density by $|\Delta\mathrm{elpd}| \leq 2.1$, far below the standard error of the difference (dse 4.8–7.5): the components are not discriminated by this approximation, consistent with the within-component absorption direction in Proposition 1. We use PSIS-LOO as a predictive comparison diagnostic (Vehtari et al., 2017; Gelman et al., 2014) rather than evidence for a unique variance attribution. Direct three-fold refits give no-interaction minus full elpd -0.5 (paired-cell SE 6.2) and no-harness minus full -6.5 (SE 3.5), with all fold diagnostics below $\widehat{R} = 1.01$ and no divergences. This supplies limited predictive discrimination, not validation of the coarse hierarchy. Removing β inflates σ_γ from 0.22 to 0.63 which shows that harness variation is absorbed by the interaction under a different within-component parameterisation. Common HalfNormal(0.5/1/2) scaling alone does not test the relative allocation between α and γ; independent-scale and row-centred analyses address that question. Across four unequal relative-scale choices—including $(s_\alpha, s_\gamma) = (2, 0.5)$ and $(0.5, 2)$—and a row-centred interaction, the posterior-mean order of all 17 model indices is unchanged from the free HN(1) reference (Kendall $\tau = 1.00$, maximum rank shift 0, top-5 overlap 5/5). The uniform-harness standardized index in Equation 4 has the same stability. This empirical stability is conditional on a narrow prior family and shared missing-edge shrinkage, and establishes neither structural robustness nor likelihood identification. Specifications and diagnostics appear in Appendix F. Restricting the fit to the largest edge–task component preserves the top five and changes the full 17-model order by only one adjacent swap (Kendall $\tau = 0.985$).

Table 8: Structural ablation and prior sensitivity (PSIS-LOO; se $\approx$ 124). Rows are independent refits under the $\widehat{R} < 1.015$ ablation gate (Section 3); the Full row is therefore not numerically identical to the Table 7 fit. All variants are predictively not distinguished by this PSIS approximation ($|\Delta\text{elpd}| \ll \text{dse}$); influential cells require the refit-based check in Appendix L. Conditional variance attribution shifts: without β, the interaction absorbs harness variance (σ_γ: 0.216 $\rightarrow$ 0.630).

Variant	elpd (se)	σ_α	σ_β	σ_γ	σ_δ	max $\widehat{R}$
Full (HN(1))	-5945.2 (124.5)	1.14	0.68	0.22	2.21	1.013
$\gamma \equiv 0$	-5944.2 (124.6)	1.13	0.72	—	2.20	1.011
$\beta \equiv 0$	-5947.3 (124.6)	1.32	—	0.63	2.21	1.004
Prior HN(0.5)	-5945.5 (124.3)	1.02	0.66	0.22	2.10	1.012
Prior HN(2.0)	-5944.5 (124.6)	1.19	0.68	0.22	2.24	1.012

B Archived Allocation Stress Test: Protocol and Trajectories

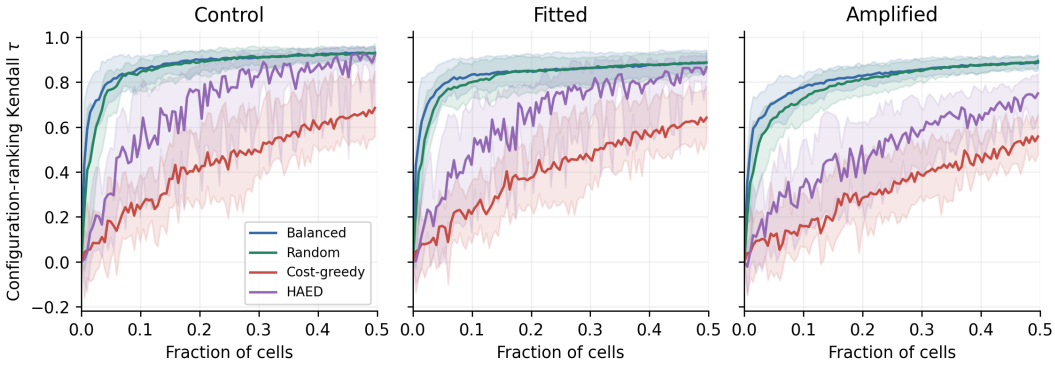

Figure 3: Configuration-ranking recovery for a HAED rule targeting α ordering. Curves show means and central 95% ranges across nine paired seeds, up to the last common observed budget (49.63%). Balanced and random allocation outperform the tested HAED proxy in fitted and amplified Gaussian regimes. Cost and latent-index curves appear in Figure 7.

This experiment predates the task-identity analysis. Its parameters are historical literal-label working-model scales, not validated estimates for a version-resolved corpus. The generated task-additive panels remain well-defined statistical test cases. The acquisition optimizes α ordering, whereas the emphasized configuration-ranking metric evaluates a different objective. This mismatch and the shared approximations prevent single-mechanism attribution.

Prospective design comparisons require known ground truth. We generate semi-synthetic evaluation tensors (12 models $\times$ 6 harnesses $\times$ 60 tasks) with parameters set to the Appendix A posterior means (σ_α=1.16, σ_β=0.68, σ_δ=2.21), under three interaction regimes: *control* ($\gamma \equiv 0$, sanity), *fitted* (σ_γ=0.216), and *amplified* (σ_γ=0.65), nine paired seeds each. Per-cell costs follow a synthetic three-tier pricing scheme (frontier/mid/small models at $3\times/1.5\times/1\times$ with mild log-normal noise; Appendix I), standing in for real API pricing; the cost-axis conclusions should be read with this proxy in mind. Four allocation strategies observe cells sequentially (batch 16, 5 trials per cell, up to 50% of the grid), refitting a diagonal-Gaussian variational posterior (Blei et al., 2017) each round; we track Kendall τ of the posterior-mean model-index ranking and the 72-configuration ranking against their generative ground truth on two budget axes: cell count and cumulative cost. The historical reanalysis emphasizes configuration ranking, which is invariant to the main-effect/interaction transformation in Proposition 1. Recovery of the generated α is a secondary, coordinate-dependent diagnostic. Setting $a_m = -2\alpha_m$ reverses its target ordering while preserving all configuration logits: on the 27 archived panels the maximum numerical change is 8.9×10^{-16}. Consequently, the sign of Kendall correlation with that latent target can reverse without changing any outcome probabilities. This gauge check re-labels the target; it does not rerun acquisition under a transformed prior.

Strategies. *Uniform* (balanced grid round-robin), *Random*, *Cost-greedy*, and *HAED*: batch-greedy maximisation, with the posterior frozen within each batch, of expected reduction in summed pairwise-ordering entropy $a(u) = \sum_{i<j} H[\Pr(\alpha_i > \alpha_j \mid \mathcal{D})] - \mathbb{E}_{y_u} H[\Pr(\alpha_i > \alpha_j \mid \mathcal{D} \cup \{y_u\})]$, divided by cell cost, with a diagonal Fisher-information/bivariate-normal approximation. The implemented score evaluates binomial information at mean task difficulty ($\delta_t = 0$); candidates sharing a model–harness pair therefore receive the same base score before batch selection. It neither computes exact expected information gain nor propagates posterior covariance. Expected information gain is a conventional objective in Bayesian experimental design (Chaloner & Verdinelli, 1995; Oladyshkin & Nowak, 2019; Xu & Liao, 2020), but sequential approximations and objectives can behave differently under limited budgets (Pauwels et al., 2014; Agrell & Dahl, 2021; Kleinegesse et al., 2021). Related adaptive-testing work likewise shows that the acquisition rule and measurement hierarchy must be evaluated together (Wang & Kingston, 2019). For the uniform baseline, each 72-cell block visits every model–harness pair once while cycling through a randomly permuted task order. Each condition–seed pair also fixes one potential binomial outcome per cell, so strategies that visit the same cell share its observation noise. Random samples without replacement.

Zero-interaction control. The $\gamma \equiv 0$ regime checks whether any advantage of the acquisition rule requires model–harness interaction. No general theorem requires equality with, or inferiority to, uniform allocation in this regime because task difficulty, cost, and current posterior uncertainty remain heterogeneous. Empirically, HAED has lower budget-averaged recovery than uniform allocation for both ranking targets on both axes. Table 9 reports paired mean differences and sample SDs across seeds; Appendix H gives the trajectories.

Main result (configuration ranking). In the fitted and amplified Gaussian regimes, on both budget axes, HAED recovers the 72-configuration ranking less accurately than balanced and random allocation (Table 9). For each seed, we first average paired $\Delta\tau$ over the 0–50% cell-budget grid or the strategies' common observed cost support. Means and sample SDs are then computed across the nine paired seed-level effects. The corresponding paired-seed bootstrap intervals are retained in Table 10 for effect-direction checks. The model-index effects are secondary simulator-coordinate diagnostics. Comparisons with random are also negative; comparisons with cost-greedy are positive. Every informative-regime cell-axis comparison with balanced uniform is negative in all 9/9 seeds (Figure 3).

Table 9: Archived HAED minus balanced-uniform ranking recovery: mean ± sample SD of budget-averaged paired Kendall-τ differences across nine seeds. SD uses denominator eight and describes seed variability. Cost effects use common observed support without endpoint extrapolation.

Ranking target	Regime	Cell axis	Cost axis
Model index	Control	-0.123 ± 0.069	-0.093 ± 0.073
	Fitted	-0.144 ± 0.055	-0.123 ± 0.057
	Amplified	-0.200 ± 0.075	-0.185 ± 0.069
Configuration	Control	-0.177 ± 0.062	-0.144 ± 0.062
	Fitted	-0.190 ± 0.060	-0.164 ± 0.063
	Amplified	-0.312 ± 0.056	-0.278 ± 0.051

Inference-approximation check. The sequential loop scores candidates with a diagonal-Gaussian variational posterior, which may contribute to HAED's deficit. We replayed HAED's trajectories in the fitted regime (3 seeds) and re-fitted the observed cells with full NUTS at 10%, 25%, and 50% budget (9 checkpoints, short 4×1000 chains, max $\widehat{R} = 1.020$; this diagnostic exceeds the main-fit threshold, so the check is directional). For model-index ranking, mean τ is 0.785 under NUTS, 0.761 under the variational loop, and 0.892 under balanced allocation. For configuration ranking, the means are 0.747, 0.669, and 0.849. NUTS improves HAED's variational estimate, especially for configuration ranking, but remains at or below balanced allocation at all 9/9 checkpoints for both targets. This directional replay suggests an inference contribution but does not isolate it: acquisition is not rerun with NUTS, the comparator is not refitted symmetrically, and the chains miss the main diagnostic gate.

Mechanism variants. No tested one-factor change reverses the deficit (Figure 4). On the fitted cell axis, removing cost division or adding a coverage constraint changes mean $\Delta\tau$ from -0.172 to

-0.128 or -0.149, whereas the diversity penalty worsens it to -0.196. Under amplified interactions, the original and coverage variants are nearly identical (-0.168), while no-cost and diversity are worse (-0.192 and -0.196). Cost-axis results are likewise regime-dependent. The variants therefore support no stable single-mechanism attribution. In this tested setting, balanced allocation remains the stronger baseline for estimating twelve model indices; the task-agnostic information proxy and missing covariance remain unisolated contributors.

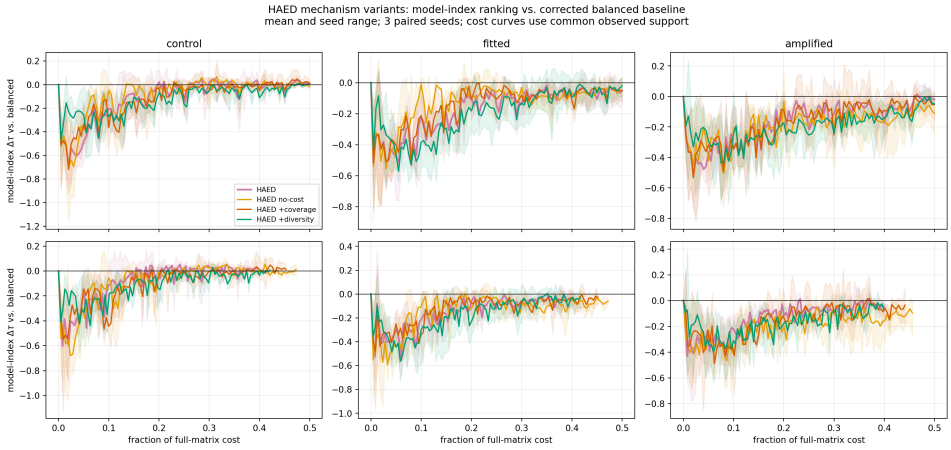

Figure 4: HAED mechanism variants: paired model-index $\Delta\tau$ relative to balanced uniform allocation (3 paired seeds; mean and seed range). No one-factor change reverses the deficit, and the direction of any improvement varies by interaction regime and budget axis. Cost curves use common observed support.

A more informative acquisition diagnostic. A separate small-panel experiment uses the same Laplace inference for all policies and a task-aware, full-covariance contrast-variance score, with uniform-harness $C_m(q)$ and held-out probabilities as invariant targets (Appendix M). Across four regimes and 12 paired seeds, its ranking-difference intervals relative to balanced allocation include zero; it does not reproduce HAED's consistent deficit. This is a different inference model and utility, not a direct replacement of HAED. Together the experiments support testing a specific acquisition implementation against strong baselines, rather than a general preference for balanced over active design. The original HAED result reverses the direction anticipated in the dated internal analysis plan.

Table 10: Paired-seed bootstrap uncertainty for the archived HAED minus balanced-uniform mean effects. Each row bootstraps nine budget-averaged paired seed differences; the 95% CI assesses the direction of the mean effect. Table 9 reports the corresponding seed dispersion. Cost effects use common observed support.

Ranking target	Regime	Axis	Mean difference	95% bootstrap CI
Model index	Control	Cell	-0.123	$[-0.166, -0.080]$
		Cost	-0.093	$[-0.138, -0.048]$
	Fitted	Cell	-0.144	$[-0.177, -0.109]$
		Cost	-0.123	$[-0.156, -0.087]$
	Amplified	Cell	-0.200	$[-0.249, -0.157]$
		Cost	-0.185	$[-0.230, -0.144]$
Configuration	Control	Cell	-0.177	$[-0.214, -0.137]$
		Cost	-0.144	$[-0.181, -0.105]$
	Fitted	Cell	-0.190	$[-0.226, -0.151]$
		Cost	-0.164	$[-0.201, -0.124]$
	Amplified	Cell	-0.312	$[-0.346, -0.277]$
		Cost	-0.278	$[-0.309, -0.247]$

C Proofs

Proposition 1. Write the linear predictor of equation 1 as $\eta = X\theta$ with $\theta = (\mu, \alpha, \beta, \gamma, \delta)$. We proceed in two stages.

Stage 1 (configuration–task contrasts). Absorb the edge-level terms: $\theta_{mh} = \mu + \alpha_m + \beta_h + \gamma_{mh}$, so $\eta_{mht} = \theta_{mh} - \delta_t$. A null direction (u, v) of this sub-design satisfies $u_e = v_t$ for every observed incidence (e, t), where $e = (m, h)$. Equality therefore propagates along every path in the bipartite edge–task graph $\mathcal{B}$. Within each connected component C, $u_e = v_t = c_C$; conversely, adding c_C to all θ_e and δ_t in C leaves every predictor unchanged. Thus the null space has exactly one location constant per component, and all within-component contrasts of θ_e and of δ_t are likelihood-identified. No minimum degree assumption is required: a leaf edge is identified relative to the task to which it is attached, whereas an isolated observed edge–task pair forms its own component.

Stage 2 (main effects and interaction). The decomposition of θ is not identified in the free-interaction model. For any model-specific constants a_m, the transformation

$$\alpha'_m = \alpha_m + a_m, \qquad \gamma'_{mh} = \gamma_{mh} - a_m$$

leaves every observed linear predictor unchanged; analogous harness-specific shifts can be absorbed into γ. Hence $\alpha_i - \alpha_j$ is not a likelihood-identified contrast, regardless of graph connectivity. If both $\mathcal{B}$ and $\mathcal{G}$ are connected and $\gamma \equiv 0$, standard location constraints identify global additive model and harness contrasts. With disconnected $\mathcal{B}$, the exact condition is full rank of the constrained additive design because the restriction can couple otherwise separate edge–task components. Other declared constraints can define different estimands, but their interpretation must be reported. The fitted α_m values in this paper are therefore prior-regularised latent indices. □

Proposition 2. In the saturated edge parameterisation, for configurations in distinct components of $\mathcal{B}$, the likelihood is invariant to independent shifts of their component-specific locations. With free interactions, these shifts are absorbed by the corresponding γ_{mh} terms. A cross-component posterior contrast consequently depends on the chosen hierarchical prior, shared hyperparameters, and identifying constraints. Under the centred exchangeable prior used here, a conditional prior calculation may be reported as a reference, but it is not a universal normal law or a prior-free structural lower bound. A bridge edge would add direct information. This is why cross-component comparisons should be accompanied by prior and constraint sensitivity analyses. □

D Constructive Free-Interaction Diagnostic

To make the absorption direction in Proposition 1 executable, we construct a $12 \times 6 \times 60$ binary panel with five trials per cell and apply three nonzero model-specific shift vectors a_m. For each shift, we set $\alpha'_m = \alpha_m + a_m$ and $\gamma'_{mh} = \gamma_{mh} - a_m$. Across all 4,320 cell predictors, the largest numerical change is at most 1.78×10^{-15} and the binomial log-likelihood difference is exactly zero at displayed precision. Yet the induced α ordering changes at 6–10 positions, while the Gaussian prior penalty changes by 7.69–59.20 under unit scales and by 34.86–209.26 under $(\sigma_\alpha, \sigma_\gamma) = (2, 0.5)$. This deterministic construction is not a posterior sensitivity analysis or a new empirical result; it demonstrates that a prior or explicit constraint selects the decomposition among likelihood-equivalent parameterisations. The script and machine-readable output are included in the anonymous supplementary archive.

E Ablation Details

Table 8 reports the archived PSIS-LOO comparisons for the structural variants and prior scales. The omitted unit is a complete binomial (m, h, t) cell, including all its repeated trials, not one individual trial. This targets prediction of an unobserved evaluation cell among represented entities, not prediction of a new model, harness, or benchmark. A fresh diagnostic of the main 16,000-draw posterior flags 27/4,702 cells (0.574%) with $\hat{k} > 0.7$, with maximum $\hat{k} = 1.378$. Per-model and per-harness fractions and maxima are released. Small elpd differences relative to their standard errors do not establish robustness to importance-sampling failure. We therefore treat Table 8 as approximate non-discrimination and supplement it with refit-based three-fold prediction in Appendix L.

F Relative-Prior and Constraint Sensitivity

The free-interaction fits below vary the HalfNormal hyperprior scales on σ_α and σ_γ independently; the row-centred fit instead imposes $\sum_h \gamma_{mh} = 0$ for every model. Each fit uses four chains with 2,000 warmup and 2,000 retained draws per chain. The free HN(1) reference and two variants have maximum $\widehat{R}$ between 1.010 and 1.013, within the predeclared ablation gate of 1.015 but above the stricter 1.01 main-fit gate. The row-centred fit has maximum $\widehat{R} = 1.008$. Across these specifications, both the prior-regularised α order and the uniform-harness standardized order are unchanged (Table 11). A separate no-interaction fit has Kendall $\tau = 0.985$ and maximum rank shift 1, but its short-run maximum $\widehat{R} = 1.033$ fails the ablation gate and is not used as primary sensitivity evidence. Predictive elpd differences among the displayed fits are at most 2.6, compared with an elpd standard error of about 124.

We also refit only the largest component of $\mathcal{B}$ (17 models, 26 harnesses, 89 tasks, 4,266 cells, 20,379 trials) with four chains, 3,000 warmup iterations, and 3,000 retained draws. The fit passes the main diagnostic gate (maximum $\widehat{R} = 1.009$, minimum ESS 609). Relative to the full-data free HN(1) fit, the 17-model order has Kendall $\tau = 0.985$, maximum rank shift 1, and top-5 overlap 5/5. Its posterior means are $\sigma_\alpha = 1.141$, $\sigma_\beta = 0.634$, $\sigma_\gamma = 0.253$, and $\sigma_\delta = 2.130$. Thus excluding the smaller literal-label component changes the archived ordering little. This is an identity-policy-specific subset analysis, not a validation of task content or of the candidate namespace joins.

Table 11: Independent hyperprior-scale and interaction-constraint sensitivity. Rank comparisons use posterior means relative to the free HN(1) fit. The same statistics are obtained for $C_m(q)$ with uniform q.

Interaction	s_α	s_γ	elpd	max $\widehat{R}$	Kendall τ	max shift
Free	1.0	1.0	-5943.8	1.013	1.000	0
Free	2.0	0.5	-5943.8	1.012	1.000	0
Free	0.5	2.0	-5942.9	1.007	1.000	0
Free	1.0	0.5	-5941.4	1.010	1.000	0
Free	1.0	2.0	-5943.6	1.007	1.000	0
Row-centred	1.0	1.0	-5944.0	1.008	1.000	0

G Baseline Index Display

H Sanity Check

I Variant Ablation Protocol

All variants share the Fisher-information proxy for expected pair-entropy reduction and differ in one mechanism each: *haed-nocost* removes the $1/c_u$ division; *haed-cov* restricts candidates to models whose observed cell count is within $+30$ of the minimum; *haed-div* multiplies the score of the k-th within-batch pick of the same model by $1/(1+k)$. 27 runs (3 variants $\times$ 3 regimes $\times$ 3 seeds). Each variant is compared with the corrected balanced trajectory on both cell and common-support cost axes. Results in Figure 4.

J Data Provenance

Terminal-Bench 2.0 contributes 32,803 normalized trial records; SWE-bench five contributes 323 aggregated records; METR Time Horizon 1.1 contributes 5,207 model–task rows, for 38,333 records in the cross-source inventory. Within the 27,948 otherwise eligible single-model Terminal-Bench trials, 8,345 have no raw model identifier, 20,415 have empty or brace-only harness-detail metadata, and 1,326 use the three curated aliases (441 `aurora-01-21`, 442 `deepseek-chat`, and 443 `opus`). These flags overlap. After alias normalisation and restriction to the dominant model–harness

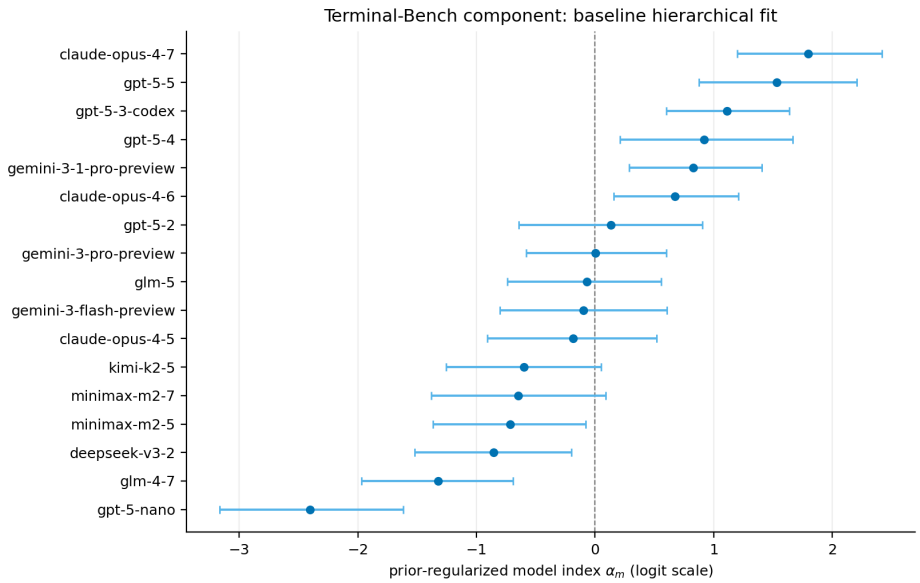

Figure 5: Posterior means and 90% credible intervals for the baseline prior-regularised model index α_m. The ordering is conditional on the free-interaction hierarchy and should not be read as a harness-invariant ranking of intrinsic model capability. The observed-supported crossover screen is Figure 2 in the main text.

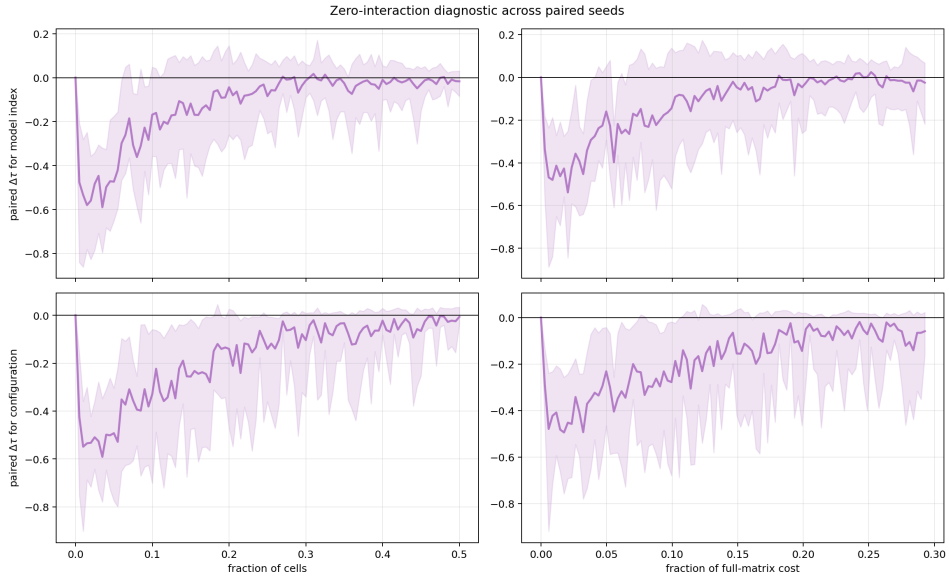

Figure 6: Sanity check under the control regime ($\gamma \equiv 0$): paired $\Delta\tau$ (HAED − balanced uniform) across nine seeds on both budget axes. The top row targets the model index and the bottom row the configuration ranking. This regime diagnoses whether the acquisition rule's behavior requires model–harness interaction; no equality is theoretically required. Mean effects are $-0.123/-0.093$ for the model index and $-0.177/-0.144$ for configuration ranking on the cell/cost axes; all paired-seed bootstrap intervals exclude zero.

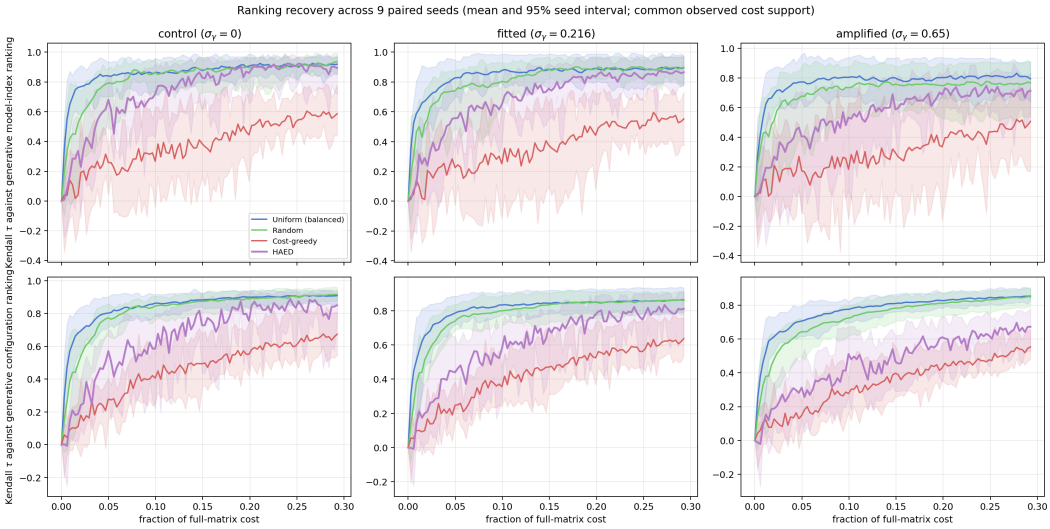

Figure 7: Budget–τ curves on the cumulative-cost axis (mean and central 95% seed interval, 9 paired seeds), restricted to common observed support with no endpoint extrapolation. The top row targets the model index and the bottom row the configuration ranking. Balanced uniform and random allocation have higher average recovery than HAED; cost-greedy trails. Exact comparator-specific supports are released.

component, the fit contains 22,818 trials in 4,702 observed model–harness–task cells. A conservative complete-metadata filter yields a different 8-model, 5-harness, 89-task component with 4,998 trials. Its new sensitivity fit describes a changed population; it does not validate missing identities. Normalisation rules, the record-level identity map, and the structured report ship with the dataset.

Table 12: Components under the archived literal task-label policy. No literal label links the two rows. Candidate prefix normalization connects the graph (Table 1); task-version equivalence has not been established. These rows are not verified disjoint task sets.

Component	Configuration edges	Task labels	Observed cells	Trials
1	48	89	4,266	20,379
2	5	90	436	2,439
Total	53	179	4,702	22,818

K SUPPORT-AWARE STANDARDIZATION AND CROSSOVER CERTIFICATES

The numerical summaries in this appendix and Appendix L reuse the historical literal-label fit. The label-policy and provenance analysis in Section 6 qualifies their task and component interpretation.

Identification of a standardized contrast. Under the free-interaction model, consider a linear contrast $L = \sum_{e \in \mathcal{E}_O} w_e \theta_e$. It is likelihood-identified if and only if $\sum_{e \in C} w_e = 0$ for each connected component C of $\mathcal{B}$. This follows immediately from the null-space characterization: the shift of L under component locations (c_C) is $\sum_C c_C \sum_{e \in C} w_e$. It vanishes for every (c_C) exactly under the stated condition. A nonzero coefficient on an unobserved edge introduces another free direction. For $C_i(q) - C_j(q)$, both models therefore need observed edges where $q_h > 0$, and their q-mass must match within every edge–task component. Specifying q defines the target. The component-mass condition establishes its identification in the fixed design under the task-additive predictor. Model fit, finite-sample precision, logistic separation, and causal interpretation require their corresponding diagnostics and assumptions.

Weighted centering as an estimand choice. Let $\bar{\gamma}_m(q) = \sum_h q_h \gamma_{mh}$. The transformation $\alpha_m^{(q)} = \alpha_m + \bar{\gamma}_m(q)$ and $\gamma_{mh}^{(q)} = \gamma_{mh} - \bar{\gamma}_m(q)$ imposes $\sum_h q_h \gamma_{mh}^{(q)} = 0$ while preserving every θ_{mh}. Consequently, $C_m(q) = \mu + \alpha_m^{(q)} + \sum_h q_h \beta_h$. If the prior is pushed forward with this coordinate change, the posterior of $C_m(q)$ is identical draw by draw. Replacing the induced prior with new independent priors is instead a new statistical model, as in the earlier row-centered sensitivity fit. Model-specific observed-support centering replaces q with q_m^O and changes the deployment mixture. It cannot create a shared harness environment where none is observed.

Observed support. Table 13 gives support for uniform q over 30 harnesses and for a common empirical q proportional to the number of fitted trials on each harness. For uniform q, the conditional effective number of observed harnesses equals the observed degree. The supplementary CSV also reports effective support for empirical q and 90% posterior intervals for own-support indices. No harness lies in $\bigcap_m O_m$. Hence there is no nonempty common observed-support reference distribution for the complete model set. Pair-specific intersections are possible but change the reference population across pairs and do not, in general, induce one transitive global ranking. Own-support ranking has Kendall $\tau = 0.897$ relative to uniform standardization and a maximum shift of two positions. This is estimand sensitivity, not validation of either global order. Component-location assumptions still apply to these posterior summaries.

Table 13: Observed support and conditional posterior-mean ranks. Uniform support mass is degree/30; empirical q weights harnesses by trial frequency. Own-support ranks use different harness mixtures across models and are not a shared-environment ranking.

Model	Degree	Uniform (%)	Empirical (%)	Uniform rank	Own rank
claude-opus-4-5	2	6.7	5.1	11	10
claude-opus-4-6	12	40.0	54.8	6	6
claude-opus-4-7	4	13.3	12.6	1	1
deepseek-v3-2	1	3.3	13.3	15	15
gemini-3-1-pro-preview	5	16.7	21.3	5	4
gemini-3-flash-preview	2	6.7	5.8	10	9
gemini-3-pro-preview	3	10.0	13.6	8	7
glm-4-7	2	6.7	15.3	16	16
glm-5	2	6.7	15.3	9	11
gpt-5-2	1	3.3	4.7	7	8
gpt-5-3-codex	9	30.0	42.4	3	5
gpt-5-4	1	3.3	5.8	4	2
gpt-5-5	4	13.3	9.7	2	3
gpt-5-nano	1	3.3	3.8	17	17
kimi-k2-5	1	3.3	13.3	12	13
minimax-m2-5	2	6.7	15.3	14	14
minimax-m2-7	1	3.3	3.1	13	12

Separating observed evidence from missing-edge predictions. The original maximum ranges over $\binom{30}{2} = 435$ harness pairs for every model pair. Only four model pairs have at least two jointly

observed harnesses. Their admissible counts are 1, 3, 3, and 1 (Table 14). All corresponding within-harness contrasts are connected in $\mathcal{B}$. For the remaining 132 pairs an observed-only maximum is undefined, not zero. The one observed-supported maximum above 0.50 concerns Claude Opus 4.6 versus Gemini 3.1 Pro Preview (0.739). The second full-screen exceedance requires an unobserved configuration. These posterior probabilities remain conditional on the inadequate baseline hierarchy and on maximizing over the specified candidate set. Neither the four-pair table nor the archived heatmap is a calibrated detection procedure. In particular, posterior Monte Carlo precision is not uncertainty calibration for selecting a maximum. A confirmatory study would freeze the pair, harnesses, task set, versions, and error criterion before collecting new trials.

Table 14: Only model pairs admitting an observed-only crossover screen. Each all-harness maximum uses 435 harness pairs; the observed-only count is pair-specific. All four observed-only cases also satisfy within-harness component connectivity.

Model pair	Observed pair count	Observed max	All-harness max
Opus 4.6 / Opus 4.7	1	0.000	0.034
Opus 4.6 / Gemini 3.1 Pro	3	0.739	0.739
Opus 4.6 / GPT-5.3 Codex	3	0.083	0.235
Gemini 3 Pro / GPT-5.3 Codex	1	0.000	0.023

L Predictive Adequacy, Structural Sensitivity, and Refit-Based Validation

All reported fits and folds retain 179 literal task labels except the explicitly restricted subset. They have not been recomputed under the candidate identity joins. In particular, task-block standard errors cluster literal labels rather than reconciled task versions.

Replicated-data checks. We reuse 2,000 evenly spaced draws from the archived 16,000-draw main posterior and independently generate $y_c^{\mathrm{rep},s} \sim \mathrm{Binomial}(n_c, p_c^s)$ for each observed cell c, holding the observed design and trial counts fixed (seed 20260906). The Pearson discrepancy is

$$T(y, p^s) = \sum_c \frac{(y_c - n_c p_c^s)^2}{n_c p_c^s (1 - p_c^s)}.$$

We compare $T(y^{\mathrm{rep},s}, p^s)$ with $T(y, p^s)$ draw by draw. The medians are 4,679 and 12,872 respectively, with zero upper-tail exceedances among 2,000 draws. This is a posterior predictive diagnostic, while frequentist significance requires a separate sampling design. Marginal 90% replicated-cell intervals cover 92.15% of observed cell counts. Their discreteness and extensive 0/1 outcomes make coverage alone an insensitive fit check.

All aggregate rates for 17 models, 30 harnesses, 179 tasks, two components, and eight distinct trial-count groups fall within their marginal 90% predictive intervals. Available submission/date fields subdivide some binomial cells. For these checks we preserve the number of trials per metadata stratum and simulate those counts using the same cell probabilities. Two of 55 submission aggregates and five of 78 date aggregates fall outside the intervals. These descriptive, overlapping checks localize discrepancies, while temporal drift and conditional dependence require targeted tests. All stratum-level observations, predictive intervals, and tail frequencies are released.

Additional sensitivity fits. These analyses extend the dated plan in response to review. Each uses four NUTS chains, where chain lengths, rank-normalized $\widehat{R}$, effective sample sizes, divergences, and all scales are reported in the released summaries. The model–task extension adds $\zeta_{mt} \sim \mathcal{N}(0, \sigma_{mt}^2)$ with $\sigma_{mt} \sim \mathrm{HalfNormal}(1)$ to Equation 1. An observed-edge-only fit uses $\theta_e = \mu + \sigma_e z_e$, $z_e \sim \mathcal{N}(0,1)$, $\sigma_e \sim \mathrm{HalfNormal}(1)$ for the 53 observed configurations only, with the original task prior. Its parameter set contains the 53 observed configurations, while cross-component locations remain prior-regularized.

The strict-provenance fit excludes missing raw model identifiers, the three curated aliases, and empty harness details, then selects the largest remaining model–harness component. This defines the 4,998-trial complete-metadata population. A separate identity-split analysis replaces each canonical model

label with its canonical-parent/raw-label pair. Missing raw labels are split by submission. This conservative partition preserves reported version distinctions and may over-split equivalent names. The resulting groups define a sensitivity population distinct from the original 17 entities. Missing prompts, tools, and harness revisions remain outside this partition.

Table 15: Post-review structural and provenance sensitivity. Dispersion ratio divides median observed by median replicated Pearson discrepancy. The last column is the replicated upper-tail frequency. Ratios across likelihoods use their own conditional variances and are diagnostics, not common predictive scores. † misses the main diagnostic gate and is not used as validated sensitivity evidence.

Fit	Models	Trials	Max $\widehat{R}$	Div.	Dispersion ratio	Tail
Model–task	17	22,818	1.0046	0	1.69	0.0075
Observed edges only	17	22,818	1.0055	0	2.77	0.0000
Strict metadata	8	4,998	1.0043	0	1.86	0.0050
Identity split†	44	22,818	1.0136	0	2.70	0.0000

Table 16: Conditional posterior uncertainty for the model–task extension. The estimate and 90% posterior credible interval (CrI) use the same literal-label population and hierarchy as the extension diagnostics.

Parameter	Posterior estimate	90% CrI
Model–task scale σ_{mt}	1.598	$[1.496, 1.703]$

The model–task extension estimates $\sigma_{mt} = 1.598$ (Table 16), changes the model-index order ($\tau = 0.941$), and reduces the observed/replicated median discrepancy from $12{,}872/4{,}679$ to $7{,}715/4{,}561$, while the upper-tail frequency remains 0.0075. It reveals an additional source of heterogeneity while residual dispersion remains. Its PSIS diagnostics flag 516/4,702 cells (10.97%, maximum $k = 1.541$). The direct refit comparison below therefore focuses on the three original structures.

The strict-metadata subset preserves the relative order of its eight baseline model indices ($\tau = 1$), but its discrepancy remains $2{,}102/1{,}128$ with upper-tail frequency 0.005. The observed-edge-only model similarly has discrepancy $12{,}962/4{,}682$ with zero exceedances. Together these checks localize the baseline discrepancy beyond metadata filtering and missing-edge prediction.

The identity partition produces 44 reported-label groups and 4,878 cells from the same 22,818 trials. Its initial fit had 24 divergences. Increasing target acceptance to 0.99 and using 4,000 warmup iterations plus 3,000 draws per chain removes divergences, while maximum $\widehat{R} = 1.0136$ remains above the main 1.01 gate (minimum bulk ESS 541). Its $13{,}107/4{,}860$ discrepancy with zero exceedances is reported as a convergence-qualified sensitivity. The version-resolved analysis awaits direct identity evidence.

Prediction unit and robust comparison. We additionally refit the full, no-interaction, and no-harness hierarchies under a shared three-fold partition of the 4,702 observed cells. Sorted cells are randomly permuted with seed 20260906 and assigned round-robin to folds. All repeated trials of one cell remain together. Held-out log predictive densities are computed by log-mean-exp over the training posterior, integrating parameter uncertainty without importance weights. This estimates prediction for omitted cells within the observed public design. Transfer to new entity versions or unseen benchmarks defines a separate prediction target. Relative results and their paired cell-wise standard errors are reported below. Because cells may remain dependent within tasks or submissions, those standard errors summarize variation at the stated cell unit. We also report a task-block sensitivity SE: sum paired cell differences within each of the 179 tasks, then compute $\sqrt{179\,\mathrm{Var}(d_t)}$ using sample variance. Independent version and submission clusters require corresponding data.

Deleting interaction changes total held-out elpd by -0.46 (cell SE 6.23; task-block SE 6.25). Deleting the harness main effect changes it by -6.51 (cell SE 3.46; task-block SE 3.62). This modest loss supplies limited evidence for a unique harness main-effect allocation. Across all nine retained

Table 17: Three-fold held-out-cell log predictive density from direct refits. Difference is relative to Full. Task-block SE aggregates paired differences over 179 tasks; neither uncertainty summary covers unrecorded version or submission dependence.

Variant	Held-out elpd	Difference	Cell SE	Task SE	Max $\widehat{R}$	Div.
Full	-5913.9	+0.0	0.0	0.0	1.009	0
$\gamma = 0$	-5914.3	-0.5	6.2	6.3	1.008	0
$\beta = 0$	-5920.4	-6.5	3.5	3.6	1.008	0

refits, maximum $\widehat{R}$ is 1.0094 and there are no divergences. Initial fits missing the diagnostic gate were lengthened before these predictive results were computed.

M Task-Aware, Covariance-Aware Small-Panel Design Diagnostic

Here the task-additive data-generating model makes $C_m(q)$ an invariant contrast target. Unrestricted model–task effects define the alternative identification problem in Equation 5.

The archived large-panel cell-axis summaries use a 101-point 0–50% grid, carrying the final 49.63% checkpoint to 50% over its last short interval. The revised primary figure stops at observed support, and archived cost comparisons already use that same support without endpoint extension.

To test whether a negative result extends beyond the original proxy, we run a separate $6 \times 4 \times 20$ panel experiment. All allocation rules use the same fixed-prior Laplace posterior over the 24 configuration logits and 20 task difficulties. The configuration prior covariance is $K_{ee'} = 4 + 1.16^2 \mathbf{1}\{m = m'\} + 0.68^2 \mathbf{1}\{h = h'\} + 0.216^2 \mathbf{1}\{e = e'\}$, and the task prior is independent with variance 2.21^2. These hyperparameters are fixed across conditions and strategies. They are not learned from the held-out outcomes. The target is the pairwise contrast vector Lx for $C_m(q)$ with uniform q over four harnesses.

For posterior mode $\hat{x}$ and full covariance V, the task-aware Fisher/Laplace score is

$$A(u) = \frac{w_u \|LVx_u\|_2^2}{c_u(1 + w_u x_u^\top V x_u)}, \qquad w_u = n_u \hat{p}_u(1 - \hat{p}_u), \quad \hat{p}_u = \mathrm{logit}^{-1}(x_u^\top \hat{x}).$$

The design vector x_u includes the candidate's configuration and task entries. The score uses a one-step linearized contrast-variance reduction from the Sherman–Morrison update. Its diagonal control replaces V by its diagonal and evaluates information with the task effect set to zero. Both controls use the same invariant utility and inference, while the archived HAED uses a different objective and approximation.

We compare balanced, random, diagonal-proxy, and task/covariance-aware allocation on 12 paired seeds per condition. Control sets $\gamma = 0$; fitted uses Gaussian interactions of standard deviation 0.216; sparse uses a 20% nonzero Gaussian mixture with marginal standard deviation 0.65; heterogeneous-trial panels use fitted interactions and $n_c \in \{1, 5, 10\}$ sampled uniformly. The other panels use five trials per cell. Costs are model-tier proxies (two models at each of 3, 1.5, and 1) times log-normal noise of log standard deviation 0.1, priced per cell even when trial counts differ. All policies share a potential binomial outcome for each cell, 24 initial cells covering every configuration, and an 80% candidate pool. A fixed 20% task holdout is never acquired. Batches contain eight cells, and budget ranges from 24 to 240 cells (5–50% of the complete panel). Posterior and acquisition scores are frozen within a batch. Balanced ordering cycles configurations within a common random task order.

We evaluate standardized ranking, configuration ranking, and probability RMSE against the known latent probabilities on the fixed held-out tasks. For each seed, we average paired differences across the 28 cell-count checkpoints, then bootstrap 12 seeds (10,000 resamples). This secondary experiment reports equal-cell-budget comparisons, with the separate cost axis in the released results. All 5,376 optimizations terminate successfully, with maximum gradient component 7.2×10^{-4}.

Table 18: Task/covariance-aware minus balanced allocation in the small-panel diagnostic (12 paired seeds). Mean differences and 95% paired-seed bootstrap confidence intervals (CIs) are separate columns; the CIs assess effect direction. Positive ranking differences and negative probability-RMSE differences favor the active rule. Comparisons use equal cell budgets and a different inference model from HAED.

Regime	Metric	Mean difference	95% bootstrap CI
Control	Standardized $\Delta\tau$	-0.020	$[-0.072, +0.030]$
	Configuration $\Delta\tau$	-0.003	$[-0.014, +0.008]$
	Probability ΔRMSE	-0.019	$[-0.035, -0.005]$
Fitted	Standardized $\Delta\tau$	-0.007	$[-0.065, +0.052]$
	Configuration $\Delta\tau$	$+0.015$	$[-0.005, +0.037]$
	Probability ΔRMSE	-0.019	$[-0.045, +0.005]$
Sparse	Standardized $\Delta\tau$	$+0.022$	$[-0.034, +0.077]$
	Configuration $\Delta\tau$	$+0.016$	$[-0.017, +0.047]$
	Probability ΔRMSE	-0.007	$[-0.021, +0.004]$
Heterogeneous n	Standardized $\Delta\tau$	$+0.014$	$[-0.022, +0.052]$
	Configuration $\Delta\tau$	$+0.016$	$[-0.002, +0.036]$
	Probability ΔRMSE	-0.002	$[-0.022, +0.018]$

The configuration and standardized-ranking CIs include zero in every condition. The control probability-RMSE interval favors task/covariance-aware allocation; the other probability-RMSE intervals include zero. Thus the broad claim that balanced design defeats active allocation is not supported. Nor does this diagnostic establish the superiority of the alternative rule. It isolates a more informative acquisition approximation under a shared inference engine, while leaving exact expected utility, learned hyperparameters, realistic costs, temporal drift, and retrospective missingness for future controlled evaluation.

N Corpus Selection and External Scope

The archived assembly code filters Terminal-Bench records to labeled, single-model trials with a score and selects the model–harness component with the largest number of vertices. Selection uses graph topology rather than success rates or fitted rankings. The dated plan envisaged a broader, version-aligned multi-source analysis, so the final corpus is a retrospective feasibility sample.

The inventory distinguishes what each source can support. Terminal-Bench provides binary trial records with task labels and more than one harness. The five assembled SWE-bench tables contain 323 percentage-resolved aggregates. Their task field is an aggregate placeholder within each table. Task-adjusted cell effects require the underlying per-task outcomes, and percentage-resolved uses a different measurement scale from a binary trial. The assembled METR source has 5,207 rows, 23 model labels, 228 task labels, and one harness label (`METR-modular-public`), with success-rate outcomes. It supports within-harness performance analysis under its own measurement definitions, while between-harness variation requires additional harnesses. We therefore report the source-specific inventory and keep incompatible outcome scales separate.

The next empirical stage is validation with a version-resolved, versioned, crossed benchmark, ideally with frozen prompts, tools, retry limits, environment and dependency versions, execution date, and adaptation history.

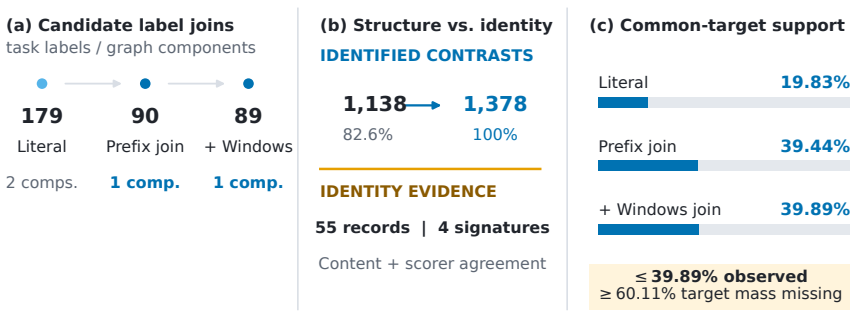

Figure 8: Connecting the incidence graph does not establish task identity or common-target support. **(a)** Candidate joins reduce 179 literal labels in two components to 90 or 89 labels in one component. **(b)** Under task additivity, identified observed-configuration contrasts increase from 1,138 (82.6%) to all 1,378; the separate provenance sample contains four reported task signatures among 55 records. **(c)** Maximum uniform harness–task support is 19.83% and 39.89% on the respective literal and joined label grids. The grid itself changes, while the 4,702 observed cells do not. Even on the joined grid at least 60.11% of target mass is unobserved. Values are deterministic design summaries, not gains from newly run evaluations.

O Executable Identity and Support Checks

Inputs and reproducibility. The identity analysis reads the released fitted-cell file. It preserves success and trial counts and rebuilds the incidence graph after each explicit candidate transformation. The input snapshot is recorded in the machine-readable summary. The released mapping contains all 179 raw task labels and their two candidate keys. Content and scorer identity fields remain unset where the source lacks the evidence required for reconciliation.

Version probe. The script pins the public dataset repository revision before listing trial files. It chooses the first lexicographic run represented in the fitted table for each submission, then the first trial directory in that run. It downloads one result per submission, yielding 55 records from the adaptive-rejection-sampler name family. This deterministic convenience sample evaluates a naming assumption. The four reported task-signature categories have counts:

Signature category	Count	Label form
A	48	bare
B	5	prefixed
C	1	prefixed
D	1	bare

Reported task configurations, repository or registry references, and source URLs appear in `task_version_evidence.json`. Credential-shaped fields were redacted in the release copy without changing the task evidence. A reported task signature is interpreted together with the generating harness, task content, and scorer.

Structural certificates. For each of the 136 model pairs and three policies, the code checks the uniform-q contrast's observed edges and per-component coefficient sums. Every uniform full-harness contrast has unsupported edges. Pair-specific intersections can support other targets, but changing that intersection between pairs does not produce one common-environment global order. Identity validity and sampling adequacy are separate from this algebraic certificate.

Support-bound checks. The released joint-support table uses uniform weights over 30 harnesses and the policy-specific 179-, 90-, or 89-label task grid. It includes plug-in observed contributions and support endpoints, explicitly marked as non-confidence intervals. The no-order certificate follows from support alone: if $s_m^{HT} < 1/2$, then $A_m \le s_m^{HT} < 1/2$ and $A_m + 1 - s_m^{HT} > 1/2$. Thus every model's admissible interval contains $1/2$. The construction allows missing probabilities to vary independently and imposes no task-additive or exchangeability model on them.

The executable checks cover a within-component contrast, a cross-component contrast, an unobserved edge, the model–task shift, admissible probability endpoints, preservation of trial counts, and the expected graph changes. These executable checks validate implementation of the structural calculations. Prospective evaluation addresses decision accuracy.

Prospective completion path. Before interpreting a version-resolved refit, recover trial-level task content, scorer and environment identifiers under documented hashing conventions, resolve candidate aliases, and report retained/excluded populations. Then refit on the resolved matrix and compare the baseline, model–task and overdispersion specifications on shared held-out units. A separate frozen crossed experiment should align acquisition and reporting targets, use actual costs when claiming cost efficiency, and validate pairwise decisions on independent repeats. These steps define the evidence needed for a version-resolved extension.

P Public Crossed-Panel Masking Protocol and Reanalysis

Provenance and unit of analysis. The source is `https://huggingface.co/datasets/open-agent-leaderboard/results`, downloaded on 7 September 2026. The released Parquet file and README are retained, together with a CSV export and retrieval metadata. The dataset declares CDLA-Permissive-2.0. Every published `benchmark_score` is retained; no missing aggregate is imputed during panel construction. The five model labels refer to DeepSeek-V3.2, Kimi-K2.5, GPT-5.2-2025-12-11, Claude Opus 4.5 and Gemini 3 Pro Preview through the source's provider routes. Architectures are Claude Code, OpenAI Solo, SmolAgents Code, tool calling and tool calling with shortlisting. Benchmarks are AppWorld test-normal, BrowseComp+, SWE-bench and τ^2 airline, retail and telecom. Exact identifiers are stored in the run manifest.

The 150 rows report 13,669 planned sessions, 13,652 total sessions and 12,021 completed sessions. Thus a complete *aggregate grid* does not imply that all planned task runs completed, or that all cells use identical realized task denominators. We use the source's primary scores rather than reconstructing success rates from successful or completed sessions. Inference is conditional on this published score table and does not certify version equivalence or underlying task-level uncertainty.

Protocol frozen before masking results. Main-text Table 4 reports the mean and sample standard deviation of the 200 mask-level percentages; seed-cluster bootstrap intervals remain in the released summary. The configuration file specifies seed 20260907, 200 replicate masks for each of three mechanisms, two target distributions, six observation fractions, ridge penalty 1 and tie tolerance 10^{-12}. Rows are sorted by model, architecture and benchmark. Random masks use a permutation of 150 cells. Block masks permute 25 model–architecture configurations and the six benchmarks within each block, revealing successive cells up to the exact budget; the last block can be partial. Score-dependent masks sort $4z_{mhb} + g_{mhb}$ in descending order, with independent standard Gumbel g_{mhb} and published score z_{mhb}. This simulates outcome-dependent availability. Reference scores enter this mask generator but never a predictor's fit for a missing cell. Budgets within a seed are nested, and both target distributions use identical masks. All generated masks and comparisons are included.

Estimation and decisions. For each model, available-score averaging is $\sum_{(h,b)\in O_m} w_{hb} z_{mhb} / \sum_{(h,b)\in O_m} w_{hb}$; it abstains if the denominator is zero. Ridge minimizes unweighted observed-cell squared error plus the sum of squared coefficients for all model, architecture and benchmark dummy effects; its intercept is unpenalized. Full dummy coding with this penalty is well-defined. Predictions are clipped to $[0, 1]$ for missing cells, while observed scores are retained. The completed table is averaged using the common target weights. There is no tuning on hidden scores. This specification serves as a transparent additive extrapolation baseline.

A support gate reports a baseline's nonzero sign only if both models have support at least κ, for $\kappa \in \{0, .2, .4, .6, .8, .9, .95, 1\}$. The sign certificate instead uses the pairwise lower and upper endpoints in Equation 6, substituting the fixed published z for p. A numerical tie produces abstention. All ten full-panel contrasts are nonzero for both targets. Always-abstain coverage is zero and its conditional error is undefined, not zero. We report pooled error counts divided by decisions and coverage divided by all pairwise cases. Seeds share one source panel. Seed-level metrics therefore summarize missingness-design variation, while deployment-risk inference requires independent task runs. For the released seed-cluster bootstrap summaries, we resample the 200 mask seeds with replacement

20,000 times. Coverage is recomputed as the mean seed-level decision fraction, and conditional error as the ratio of summed errors to summed reported decisions within each bootstrap sample. This cluster bootstrap preserves dependence among the ten model-pair decisions that share a mask.

Matched decision budgets. For each mask and target, let k be the number of pairs certified by the support bounds. For each predictive baseline separately, we sort the ten pairs by absolute predicted score difference, break ties by the fixed pair index, and report its top k signs. The rule uses only observed scores for its predictions and the certificate's observed-data decision count for its budget; it does not use full-table signs to choose pairs. Sharing k makes coverage identical *within every mask*, stronger than matching an average threshold. It nevertheless gives baselines a budget selected by another method, so these rows are a descriptive comparison rather than a prospective policy. The reanalysis reads the retained masks and recreates the original ridge and available-score predictions without generating new masks.

Table 19: Equal-coverage decisions at 90% observation. Each mask has 200 replicates and ten pairs. Errors are counts among the same number of decisions per method, using the fixed complete panel as reference. DB denotes the four-domain target; U6 weights the six benchmarks uniformly.

Target	Mask	Decisions	Certificate	Available	Ridge
DB	Random	1,233	0	0	0
DB	Config. block	1,237	0	0	0
DB	Score-dependent	1,155	0	0	0
U6	Random	1,271	0	0	0
U6	Config. block	1,242	0	1	1
U6	Score-dependent	1,313	0	2	0

At 80% observation, all equal-coverage error counts are zero for both targets and all three masks. The 90% exceptions on U6 are 1–2 mistakes among more than 1,200 selected decisions. These results test selective accuracy on this single panel; they do not imply a general error bound for the baselines.

Table 20: Fixed full-table reference scores under two common targets. Values aggregate the published benchmark scores.

Model label	Four-domain target	Uniform six-benchmark target
DeepSeek V3.2	0.4102	0.4623
Kimi K2.5	0.3650	0.3843
GPT-5.2	0.4055	0.4663
Claude Opus 4.5	0.6575	0.6863
Gemini 3 Pro	0.5872	0.6374

Verification and supported conclusion. The run retains 3,600 masks and 1,296,000 method/threshold/pair decisions. All 36,000 model-level containment checks pass with no masks fail. At full observation every certificate and non-abstaining baseline recovers the reference sign. Zero certificate errors verify the deterministic bounded-score calculation. Measured decision coverage depends on support, weighting and score separation. It does not calibrate task-run uncertainty. All fractions, targets, thresholds, raw decisions and seed metrics are released.

Next-cell protocol. The follow-up script reads the stored 80% and 90% masks and reveals 1, 5 or 10 additional cells. Random extra cells are sampled from the remaining indices with a seed derived from the frozen protocol seed. Weight policies rank missing cells by the declared target mass $q_h r_t$ and never inspect hidden scores when choosing an index. The published aggregate is inserted only after the choice. Deferred-weight restricts the pool to models that still participate in at least one uncertified pair, then applies the same mass ranking. Mean coverage values in Table 6 are averages over the 200 existing random-mask seeds on the four-domain target.