AGENT EVALUATION RELIABILITY: MORE TASKS WON'T (ALWAYS) FIX AN AGENT LEADERBOARD

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent evaluations are increasingly used to compare models, assess capabilities, and inform deployment decisions, yet observed scores can reflect not only the model but also the effects of the evaluation conditions such as the scaffolds or tasks. This makes reliability claim-dependent: an evaluation that reliably ranks deployed systems may not reliably rank underlying models or produce reliable absolute scores. We ask which conclusions current agent evaluations reliably support and what additional evaluation would actually improve them. Using Generalizability theory, we develop a Bayesian variance-decomposition framework for sparse, imbalanced agent leaderboards and apply it to 22 benchmarks from the Holistic Agent Leaderboard and Harbor Index. The framework separates *signal*, performance differences relevant to the intended claim, from *noise*, irrelevant variation that can still change scores or rankings. We find four practical results: (1) Reliability depends on the measurement goal. Fixed model–scaffold systems are ranked reliably ($E\rho^2 = 0.935$–0.994), while underlying-model reliability is substantially lower (0.148–0.841). (2) Scaffold choice can change evaluation conclusions. We introduce *inter-scaffold reliability*, measuring whether scaffolds preserve model rankings, and show that scaffold effects vary substantially across evaluations. (3) More tasks cannot resolve all uncertainty. Even infinitely many similarly constructed tasks improve model-ranking reliability of the dataset by at most 0.097 when uncertainty is dominated by limited scaffold coverage. (4) Pooling diverse benchmarks can improve cross-task rankings at lower cost. For rankings across diverse agentic tasks, pooling benchmarks raises projected reliability from 0.44 to 0.75 at the same task budget and can reduce projected cost by up to 83%. Evaluation design should therefore follow the intended claim: practitioners should identify what a score or ranking should mean, diagnose what limits its reliability, and spend evaluation budget on the sources of uncertainty that matter.[1]

1 INTRODUCTION

Agentic AI systems interact with software, tools, users, and environments to accomplish multi-step goals. Agent evaluations such as SWE-bench (Jimenez et al., 2023) and τ-bench (Yao et al., 2024) are increasingly used to test and rank such systems in technical reports, system cards, and policy discussions (Anthropic, 2026; OpenAI, 2026). Yet, a score on such evaluations is only useful if we know what conclusions it can reliably support: Does an observed ranking reflect model performance differences that persist beyond the particular evaluation setup? Would the same score or ranking hold under a different scaffold, a different set of tasks, or a broader range of agentic settings?

These questions are particularly important for agent evaluations because performance depends on more than the underlying model. A scaffold[2] manages the state, exposes tools, and translates model outputs into actions; together, the model and scaffold form the system. This creates an immediate measurement choice: are we trying to evaluate the complete system as deployed, or isolate performance differences attributable to the underlying model? The answer changes what should count as signal and what should count as evaluation error.

[1] Code and data will be made public.

[2] In the current literature, the software infrastructure and post-training methods surrounding the LLM allowing it to act is synonymous with "scaffold", "harness", and even "agent".

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.

Reliability also depends on the conclusion we want to draw. A practitioner may care whether a model ranking remains stable, whether an absolute score would remain similar, or whether performance generalizes across different agentic tasks. These are different measurement goals and need not have the same reliability. Existing agent reliability work primarily studies whether a fixed agent behaves consistently across repeated runs or prompt perturbations of the same task (Rabanser et al., 2026; Razavi et al., 2025; Yao et al., 2024). We instead ask: **what claims do current agent evaluations reliably support, and what additional evaluation would make those claims more reliable?**

We use Generalizability Theory (G-theory) (Cronbach et al., 1972; Brennan, 2001) to separate variation associated with models, tasks, scaffolds, benchmarks, and their interactions. To word it differently, our framework separates *signal*–performance differences that persist across the conditions an evaluation is meant to generalize over–from *noise*–variation that is irrelevant to the intended claim but can still change the resulting score or ranking; reliability increases when signal dominates this noise. G-theory lets us identify not only how reliable an evaluation is for a particular claim, but also what limits that reliability and whether adding more tasks, scaffolds, or benchmark coverage would help. We develop a Bayesian variance-decomposition approach for the sparse and imbalanced designs common in agent leaderboards and apply it to $30{,}859$ rollouts from 9 benchmarks in the Holistic Agent Leaderboard (HAL) (Kapoor et al., 2025) and 13 the Harbor Index (Shi et al., 2026).

Our results yield four practical findings:

- **Reliability depends on the measurement goal**. Fixed model–scaffold systems are ranked reliably $E\rho^2_{MA} \in [0.94,0.99]$, while rank reliability for the models is substantially lower $E\rho^2_{M} \in [0.148, 0.841]$.
- **Changing the scaffold can change evaluation conclusions**. We introduce *inter-scaffold reliability*, which measures whether different scaffolds produce similar model rankings when evaluating the same models on the same tasks. It ranges from 0.205 to 0.828 across benchmarks. We further show that scaffold choice can change which tasks a model solves even without improving its overall benchmark score.
- **Adding more tasks does not always make an evaluation more reliable**. Under the observed scaffold coverage, even infinitely many similarly constructed tasks would improve model-ranking reliability by at most approximately 0.10.
- **When the goal is to rank models across diverse agentic tasks, pooling diverse benchmarks can improve reliability at lower cost**. At the same task budget, projected ranking reliability rises from approximately 0.44 within one benchmark to 0.75 across the nine-benchmark battery, while reliability-aware allocation can reduce projected evaluation cost by up to 83%.

The broader lesson is that more evaluation is not automatically better evaluation. Practitioners should first specify what they want an evaluation result to mean, then identify which sources of uncertainty limit that claim, and allocate evaluation effort accordingly.

2 Related Work and Background

Agent evaluation. Agent benchmarks span software engineering (Jimenez et al., 2023), web navigation (Zhou et al., 2023; He et al., 2024), general assistance (Mialon et al., 2023), and customer service (Yao et al., 2024). Existing reliability studies primarily test whether a fixed agent behaves consistently across repeated runs, prompt perturbations, tool configurations, or environmental failures (Rabanser et al., 2026; Razavi et al., 2025; Wang et al., 2026; Kumar and Mishra, 2025). These analyses assess robustness of a particular system. We instead study the reliability of the *comparative inference*: whether the reported ordering of models or systems would persist under new tasks, scaffolds, or benchmarks. This distinction is consequential because a system can be repeatable within one harness while its rank is highly contingent on that harness.

Generalizability theory separates signal from conditional advantage. Generalizability theory (G-theory) treats evaluation conditions as measurement *facets* and decomposes score variation into their main effects and interactions (Cronbach et al., 1972; Shavelson et al., 1989; Brennan, 2001; Cronbach and Shavelson, 2004). Related work shows that AI benchmark conclusions depend on task sampling, evaluation conditions, and the model population (Madaan et al., 2024; Hardy and Kim, 2026; Hardy et al., 2026). The crucial first choice is the *object of measurement*. Model–scaffold compatibility is signal when selecting a deployable system, but error when ranking models

independently of scaffolding. Reliability therefore belongs to an object, a generalization universe, and an evaluation design, not to a benchmark alone.

For an object o and design $\mathcal{D}$, the relative generalizability coefficient and its signal-to-noise ratio are

$$E\rho_o^2(\mathcal{D}) = \frac{\sigma_o^2}{\sigma_o^2 + \sigma_\delta^2(\mathcal{D})}, \qquad \mathrm{SNR}_o(\mathcal{D}) = \frac{\sigma_o^2}{\sigma_\delta^2(\mathcal{D})} = \frac{E\rho_o^2}{1 - E\rho_o^2}. \tag{1}$$

Here σ_o^2 is the variance of scores averaged over the specified universe, and σ_δ^2 contains variation that can change relative standing. Under $X_o = Z_o + \delta_o$, with uncorrelated universe score and error, $E\rho_o^2 = \mathrm{Corr}^2(X_o, Z_o)$.

Facet main effects cancel from comparisons only when objects share the same conditions and weights. A uniformly difficult task, for example, does not change relative latent scores under a common allocation. Object-by-facet interactions do: they represent advantages that depend on the chosen task, scaffold, or benchmark. G-theory uses these components in a *Decision study* (D-study) to project reliability before collecting more data.

3 Reliability and Generalizability for Agent Leaderboards

Agent leaderboards are often sparse, imbalanced, and partially crossed: models are evaluated through different scaffolds, task counts are unequal, and many combinations are absent. In this section, we discuss several metrics to measure reliability in agent leaderboards given these common constraints. We fit a Bayesian variance decomposition to binary task outcomes, then propagate its uncertainty into reliability, design projections, and model ranks.

3.1 Observations and Generalization Universe

Let $b \in \mathcal{B}$ index benchmarks, $i \in \mathcal{I}_b$ tasks nested within benchmark b, $m \in \mathcal{M}$ models, and $a \in \mathcal{A}$ agent scaffolds. The response $y_{bima} \in \{0, 1\}$ indicates whether model m, operated through scaffold a, solves task i from benchmark b. Our primary object is the model m. Its universe score is the component expected to persist across new tasks, benchmarks, and scaffolds resembling those represented in the leaderboard. We also consider the model–scaffold pair (m, a), the relevant object when selecting a deployable system. We treat tasks as draws from benchmark-specific construction and grading processes; benchmarks as instruments drawn from a battery of contemporary agent evaluations; scaffolds as contemporary evaluation and orchestration systems; and models as the frontier-model population represented in the data. These universes delimit the inference. In particular, generalization to new benchmark-like tasks does not establish that a benchmark represents all real-world uses.

3.2 A Latent Variance Decomposition for Binary Outcomes

We use Bayesian Bernoulli–logit mixed models: $y_{bima} \sim \mathrm{Bernoulli}(p_{bima})$, $\mathrm{logit}(p_{bima}) = \eta_{bima}$. Equivalently, $y_{bima} = \mathbb{1}\{z_{bima} > 0\}$, where $z_{bima} = \eta_{bima} + \varepsilon_{bima}$ and $\varepsilon_{bima} \sim \mathrm{Logistic}(0, 1)$, giving the standard latent logistic residual variance $\pi^2/3$. Modeling the binary responses directly avoids Gaussian approximations that are particularly misleading under floor or ceiling effects. All random effects are mean-zero Gaussian and mutually independent; for example, $u_m^{(M)} \sim \mathcal{N}(0, \sigma_M^2)$. Variance components and reliability coefficients are defined on the common latent log-odds scale (see Appendix H.1). From an Item Response Theory (IRT) perspective, these are random-item, many-facet Rasch models (Wang and Wilson, 2005; Fox and Glas, 2001; Linacre and Wright, 2002).

3.2.1 Benchmark-level model and system reliability

For each benchmark b, we fit

$$\eta_{ima}^{(b)} = \beta_0^{(b)} + u_i^{(I)} + u_m^{(M)} + u_a^{(A)} + u_{im}^{(IM)} + u_{ia}^{(IA)} + u_{ma}^{(MA)}. \tag{2}$$

Variance components are benchmark-specific, with the index b suppressed for readability. The decomposition separates persistent model differences from task sensitivity, scaffold differences, and model–scaffold compatibility. Because repeated observations of the same (i, m, a) cell are rare,

the task–model–scaffold interaction is not separately identifiable from latent response variation; we denote this terminal residual variance by $\sigma^2_{IMA,e} = \pi^2/3$. For an equal-allocation design with n_i tasks and n_a scaffolds, model-ranking reliability is

$$E\rho^2_{M(b)}(n_i, n_a) = \frac{\sigma^2_M}{\sigma^2_M + \sigma^2_{IM}/n_i + \sigma^2_{MA}/n_a + \sigma^2_{IMA,e}/(n_i n_a)}. \tag{3}$$

Scaffold main effects do not enter the denominator because shifting every model equally does not change their ordering. In contrast, model–scaffold interactions are rank-relevant: they encode which models benefit from which scaffolds. When the object is the complete *model–scaffold system*,

$$E\rho^2_{MA(b)}(n_i) = \frac{\sigma^2_M + \sigma^2_A + \sigma^2_{MA}}{\sigma^2_M + \sigma^2_A + \sigma^2_{MA} + (\sigma^2_{IM} + \sigma^2_{IA} + \sigma^2_{IMA,e})/n_i}. \tag{4}$$

To measure whether scaffold choice preserves model ordering, we additionally define the inter-scaffold reliability for two independently sampled scaffolds evaluated on the same n_i tasks:

$$\rho^{(b)}_{AA'}(n_i) = \frac{\sigma^2_M + \sigma^2_{IM}/n_i}{\sigma^2_M + \sigma^2_{IM}/n_i + \sigma^2_{MA} + \sigma^2_{IMA,e}/n_i}. \tag{5}$$

This is analogous to inter-rater reliability, with scaffolds acting as alternative measurement procedures. A low value means that changing the scaffold can change which model appears strongest, even when each scaffold yields internally stable scores.

3.2.2 Persistent model differences across pooled benchmarks

Each benchmark contains too few scaffolds to estimate all scaffold-related components precisely in isolation. We therefore also fit a joint model across the nine-benchmark leaderboard:

$$\eta_{bima} = \beta_0 + u^{(B)}_b + u^{(I[B])}_{bi} + u^{(M)}_m + u^{(A)}_a + u^{(BM)}_{bm} + u^{(BA)}_{ba} + u^{(MA)}_{ma} + u^{(IM[B])}_{bim} + u^{(IA[B])}_{bia} + u^{(BMA)}_{bma}. \tag{6}$$

Items are nested within benchmarks, while models and scaffolds are crossed with benchmarks wherever supported by the observed incidence graph. Shared models and scaffolds connect benchmarks and permit partial separation of their effects. Hierarchical priors can produce estimates for weakly supported contrasts, but cannot supply empirical identification between disconnected components; connectivity and estimability diagnostics are reported in Appendix F.

The leaderboard universe score for model m is $u^{(M)}_m$, the component expected to persist across sampled benchmarks, tasks, and scaffolds. Benchmark-conditioned capability $u^{(BM)}_{bm}$ contributes to performance on benchmark b, but is not assumed to transfer to a new benchmark. For a balanced design with n_b benchmarks, n_i tasks per benchmark, and n_a scaffolds,

$$E\rho^2_M(n_b, n_i, n_a) = \frac{\sigma^2_M}{\sigma^2_M + \sigma^2_{BM}/n_b + \sigma^2_{MA}/n_a + \sigma^2_{BMA}/(n_b n_a) + \sigma^2_{IM[B]}/(n_b n_i) + \sigma^2_{BIMA,e}/(n_b n_i n_a)}. \tag{7}$$

Here $\sigma^2_{BIMA,e} = \pi^2/3$ is terminal cell-specific variation and the logistic residual (see § E.3.1 for separated variance). Equation 7 maps each design intervention to the uncertainty it can reduce: tasks average task-indexed error, benchmarks average benchmark-conditioned differences, and scaffolds average scaffold-conditioned differences. Table 1 consolidates the estimands used in the main body.

3.3 What Additional Evaluations Can Resolve

Proposition 3.1 (Facet-specific replication and reliability ceilings)**.** *For variance components, the reliability in Equations 3 and 7 is non-decreasing in each sample size. Holding n_b and n_a fixed,*

$$\lim_{n_i \to \infty} E\rho^2_{M(b)}(n_i, n_a) = \frac{\sigma^2_M}{\sigma^2_M + \frac{\sigma^2_{MA}}{n_a}}, \quad \lim_{n_i \to \infty} E\rho^2_M(n_b, n_i, n_a) = \frac{\sigma^2_M}{\sigma^2_M + \frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{BMA}}{n_b n_a}}. \tag{8}$$

Proof. Increasing n_i monotonically decreases only denominator terms indexed by tasks. Those terms converge to zero, while benchmark- and scaffold-indexed terms remain. □

Table 1: Reliability estimands. Each row applies $E\rho_o^2 = \sigma_o^2/(\sigma_o^2 + \sigma_\delta^2)$, $\text{SNR}_o = \sigma_o^2/\sigma_\delta^2$, but changes the object whose ordering should generalize. The n_f terms denote equal allocation over facet f.

Estimand	Object of inference	σ_o^2	$\sigma_\delta^2(\mathcal{D})$	Eq.
$E\rho^2_{M(b)}$, $\text{SNR}_{M(b)}$	Does benchmark b preserve model order?	σ^2_M	$\frac{\sigma^2_{IM}}{n_i} + \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{IMA,e}}{n_i n_a}$	(3)
$E\rho^2_{MA(b)}$	Does benchmark b preserve system order?	$\sigma^2_M + \sigma^2_A + \sigma^2_{MA}$	$\frac{\sigma^2_{IM} + \sigma^2_{IA} + \sigma^2_{IMA,e}}{n_i}$	(4)
$\rho^{(b)}_{AA'}$	Do scaffolds induce the same model order?	$\sigma^2_M + \frac{\sigma^2_{IM}}{n_i}$	$\sigma^2_{MA} + \frac{\sigma^2_{IMA,e}}{n_i}$	(5)
$E\rho^2_M$, SNR_M	Does the leaderboard preserve model order?	σ^2_M	$\frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{BMA}}{n_b n_a} + \frac{\sigma^2_{IM[B]}}{n_b n_i} + \frac{\sigma^2_{BIMA,e}}{n_b n_i n_a}$	(7)

Thus, arbitrarily many tasks cannot overcome model–benchmark heterogeneity or model–scaffold coupling. This result motivates evaluating diversity, rather than task count alone, as a design resource.

Corollary 3.2 (Benchmark breadth at a fixed task budget)**.** *Fix n_a and the total number $N = n_b n_i$ of tasks per scaffold. Under the balanced, exchangeable-facet model,*

$$\sigma^2_\delta = \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{BM} + \sigma^2_{BMA}/n_a}{n_b} + \frac{\sigma^2_{IM[B]} + \sigma^2_{BIMA,e}/n_a}{N}. \tag{9}$$

Thus, distributing the same task budget across more benchmarks increases reliability whenever $\sigma^2_{BM} + \sigma^2_{BMA}/n_a > 0$.

Proof. Use $n_i = N/n_b$ in Eq. 7. Only the benchmark-conditioned term changes with n_b. □

Breadth does not create additional persistent model signal; it reduces contamination by condition-specific advantages. The corollary assumes similarly informative benchmark draws, adequate crossing, and no additional benchmark setup cost. It does not imply that arbitrary new benchmarks outperform more tasks, or that benchmarks always offer greater value than scaffolds.

3.4 Bayesian Estimation and Design Studies

We use Bayesian estimation with regularizing priors because sparse crossed designs can yield unstable variance estimates, especially near floor and ceiling performance. For every posterior draw, we compute reliability, SNR, and the task-only ceilings. D-study projections therefore retain uncertainty in the variance decomposition rather than substituting point estimates. Partial pooling allows connected observations to inform common variance components while retaining uncertainty where scaffold or cross-benchmark replication is limited.

Primary D-studies use common equal allocations; analyses reproducing the observed imbalance appear in Appendix C. We measure evaluation volume in model–scaffold–task trials and use dashboard prices for dollar-cost projections. Repeated task subsampling checks whether reduced designs retain the ranking information predicted by the D-study (see § G.1).

3.5 Posterior Capability, Ranks, and Transportability

For benchmark b, the scaffold-marginalized latent capability of model m in posterior draw s is $\theta^{(s)}_{mb} = u^{(M,s)}_m + u^{(BM,s)}_{bm}$. To test whether the estimated shared model component is transportable rather than an artifact of the nine-benchmark panel, we evaluate $\widehat{\theta}_m = u^{(M)}_m$ on four contemporaneous benchmarks excluded from model estimation. Among overlapping models, we compare its association with external performance against that of the conventional in-panel mean-accuracy aggregate. This is an out-of-panel test of convergent predictive validity, not evidence of universal deployment validity nor internal construct validity. Finally, we assess sensitivity to the link function, estimation method, variance-component estimator, prior specification, and leave-one-benchmark-out refits. The latter analysis also identifies benchmarks that disproportionately contribute signal or connectivity to the leaderboard. Appendices have full computational estimation details (§ C), robustness checks and sensitivity analyses (§ E), and methodological comparisons (§ H).

Table 2: Benchmark design summary. Counts denote unique levels of each facet.

Benchmark	Models	Tasks	Agent Scaffolds	Model–Scaffold Pairs
AssistantBench	18	33	2	30
CORE-Bench Hard	34	45	3	56
GAIA	21	165	2	34
Online-Mind2Web	13	300	2	23
SciCode	17	65	3	37
ScienceAgentBench	19	102	2	25
SWE-bench Verified Mini	24	50	2	26
τ-bench Airline	23	50	3	42
USACO	13	307	2	14

4 DATA

We analyze agent rollouts from all nine benchmarks distributed via HAL Kapoor et al. (2024): AssistantBench (Yoran et al., 2024), CoreBench Hard (Siegel et al., 2024), GAIA (Mialon et al., 2023), Online-Mind2Web (Xue et al., 2025), SciCode (Tian et al., 2024), ScienceAgentBench (Chen et al., 2024), SWE-bench Verified Mini (Jimenez et al., 2023), τ-bench Airline (Yao et al., 2024), and USACO (Shi et al., 2024), covering web navigation, scientific programming tasks, multi-step and user assistance, software engineering, customer-service interaction, and competitive programming (details in Appendix A.1). From each of AssistantBench, GAIA, SciCode and SWE-bench Verified Mini, the HAL leaderboard uses a subset of tasks from the full benchmark (Kapoor et al., 2025).

HAL dataset covers 29,923 total agent rollouts across 54 models and 13 scaffolds; 9 models appear on every benchmark each of which span at least 69% of the available scaffolds. Each rollout is over a single task, model, reasoning-effort, and scaffold. Summary statistics are found in Table 2. All outcomes are task-level binary scores. The incidence structure is incomplete at every level, a common problem in leaderboards (Singh et al., 2026). One scaffold is shared across eight benchmarks; AssistantBench connects the remaining benchmark through an additional shared scaffold; all other benchmarks contain at least one additional benchmark-specific scaffold. Models and scaffolds are neither fully crossed nor evenly replicated. Nevertheless, shared models and scaffolds connect the observation graph, permitting partial separation of model, benchmark, and scaffold effects.

Additionally, to corroborate our claims and provide external convergent validity, we use two supplemental datasets. The Harbor Index dataset is a meta-benchmark containing task level scores from 29 agent benchmarks for 9 LLMs and 4 scaffolds. The second dataset consists of LLM-level scores for four benchmarks that are contemporaneous with and include LLMs specified in the HAL data. Appendices more detail for the datasets (§ A) and data connectivity analysis (§ F). These data represent current practice: because agent evaluations are expensive,many model–scaffold–benchmark combinations absent entirely.

5 RESULTS AND PRACTICAL RECOMMENDATIONS

The results identify a structural limitation of task-only scaling: current evaluations can distinguish fixed systems precisely while leaving persistent model differences unresolved. The useful response is to change the measurement design, not merely enlarge it.

5.1 RELIABILITY DEPENDS ON THE MEASUREMENT GOAL

Reliability depends on the measurement claim a practitioner wants to make. In agent evaluations, two choices are relevant here: whether the object of measurement is the underlying model or the model–scaffold system, and whether the goal is relative (rank) or absolute interpretation of scores.

Model and system reliability can differ substantially. Across the nine benchmarks, estimated system reliability (i.e., ranking model–scaffold pairs) is $E\rho^2_{MA(b)} \in [0.935, 0.994]$, whereas model reliability is $E\rho^2_{M(b)} \in [0.148, 0.841]$ (Figure 1). These are not conflicting assessments of the same score. System reliability counts scaffold differences and compatibility as signal; model reliability requires differences that persist after averaging over scaffolds. A leaderboard can therefore be reliable for ranking model-scaffold pairs but unreliable for solely ranking the underlying model.

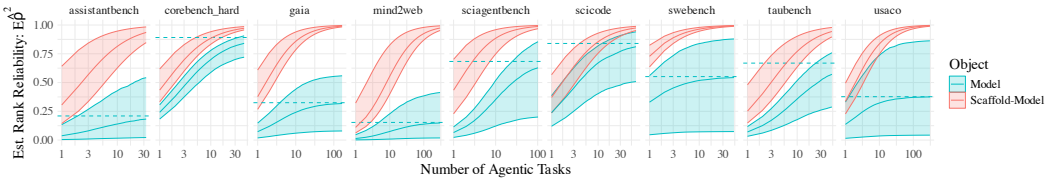


Figure 1: **Reliability depends on the object being ranked & task replication cannot remove scaffold-dependent model advantages.** D-study reliability as tasks are added for base models and fixed model–scaffold systems. Dashed lines are task-only asymptotes for model rankings. Estimates are posterior medians with 68% standard error HDIs.

Recommendation. Specify whether the intended object of measure is the model or a model–scaffold system, especially in leaderboards using multiple agentic benchmarks, and whether the intended interpretation is rank comparison (or whether absolute scores are intended to carry meaning, such as capability thresholds.[3] Provide rankings, calculate scores, and estimate reliability for that specific claim rather than reporting a single generic reliability statistic.

5.2 Changing the scaffold can change evaluation conclusions

How much the scaffold matters depends on the benchmark. Inter-scaffold reliability (Eq. 5) measures whether different scaffolds produce similar rankings or scores when evaluating the same models on the same tasks ranging from (0.205) on OnlineMind2Web to (0.828) on CORE-Bench Hard (Figure 2). At the low end, changing only the scaffold can substantially change which models appear to perform best on a benchmark, even when the models and tasks remain fixed.

Scaffolds can change which tasks get solved without improving the overall benchmark score. The pooled decomposition (Eq. 6) distinguishes persistent differences from task-specific sensitivity. In the pooled analysis, the contrast $\sigma^2_M - \sigma^2_A$ favors neither direction (posterior directional probability approximately 50%) (Makowski et al., 2019). However, task-specific variation across scaffolds is larger than task-specific variation across models with posterior probability $\Pr(\sigma^2_{IA[B]} > \sigma^2_{IM[B]} \mid y) = 84\%$ (see Figure 10). Benchmark specific decompositions (Eq. 2) show the same pattern (Figure 2, middle and right) This suggests that scaffold choice can strongly affect success on individual tasks, even without making one scaffold producing uniformly better on a benchmark as a whole. Similarly, A scaffold can change which tasks a system solves without producing a uniformly stronger system.

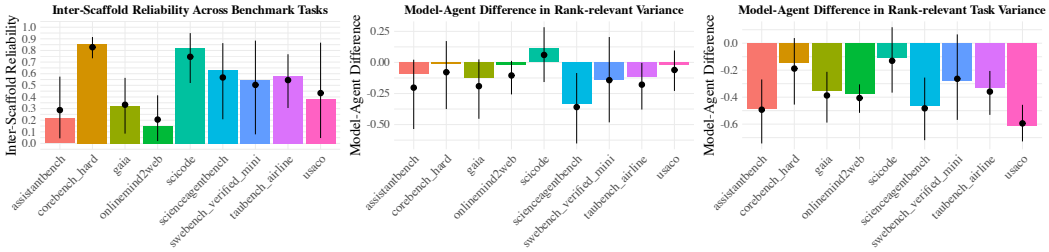


Figure 2: **Scaffolds are rank-relevant and result in distinct task-level consequences.** *Left*: posterior inter-scaffold reliability $\rho^{(b)}_{AA'}$. *Middle* and *Right*: model-minus-scaffold system rank-relevant variance contrasts (over denominator from Eq. 4) at the benchmark and task levels, respectively; positive values favor the model contribution. Task-interaction contrasts measure task-specific sensitivity. Points and bars denote posterior means and medians, respectively, with 68% standard error HDIs.

[3]There may be situations where reliability of actual numeric scores (not just ranking) is needed. We illustrate this separate estimation on the observed scale in § H.5, which shows that score reliability is substantially lower for all benchmarks. Reliable *rankings* do not imply reliable *scores*.

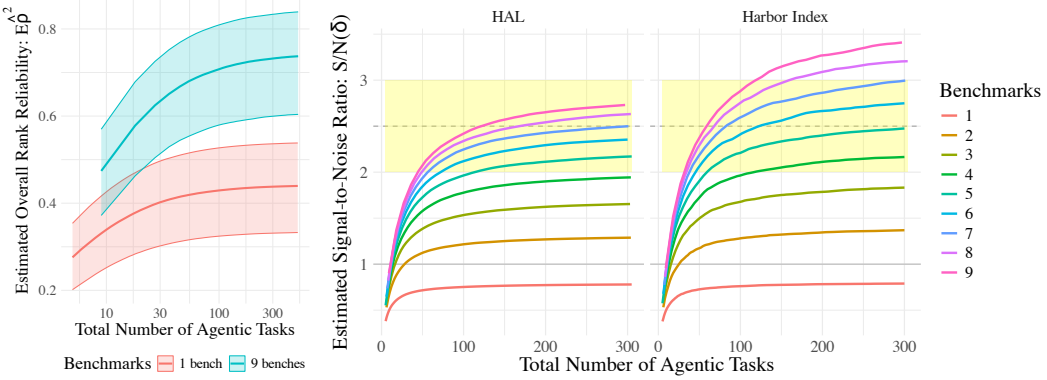
Recommendation. If the goal is to evaluate the underlying model, test the same models across multiple scaffolds rather than relying on a single implementation. Report how much conclusions change across scaffolds, and avoid attributing scaffold-specific advantages to the model itself. If the deployed model–scaffold system is the intended object of evaluation, scaffold variation can instead be treated as part of the system being measured.

5.3 More tasks cannot always resolve evaluation uncertainty

More tasks only help when the main uncertainty comes from differences across tasks. Adding tasks helps when a model's measured performance changes substantially depending on which tasks it is tested on (Proposition 3.1). It does not fix uncertainty caused by other facets, such as the choice of scaffold. In our data, even infinitely many similarly constructed tasks would improve model-ranking reliability by at most 0.10. Only CORE-Bench Hard and SciCode can exceed $E\rho^2 = 0.75$ through task scaling alone; for OnlineMind2Web, reliability increases only from 0.148 to 0.153.

A benchmark can run out of useful information before it runs out of tasks. Adding tasks repeatedly measures the same model–scaffold and benchmark-specific effects. Once these sources of uncertainty dominate, more tasks make the existing evaluation setup more precise without making the broader model claim substantially more reliable. Low reliability does not mean that a benchmark measures an unimportant capability; it means that, for the models being compared, its scores do not reliably distinguish the quantity of interest. Thus, reliability should be re-estimated as the competitor population changes.

Recommendation. Estimate how much reliability can improve from adding tasks before expanding a benchmark. Add tasks when task sampling is the main source of uncertainty; otherwise, spend evaluation budget on relevant facets that drive uncertainty, such as broader scaffold coverage.



(a) Model reliability.

(b) Rank-relevant signal-to-noise ratio.

Figure 3: **Broader measurement conditions reduce error that an increase in the number of test items alone cannot.** Pooled D-studies compare concentration within one benchmark against allocation across multiple benchmarks. Estimates are posterior medians; panel (a) includes 68% HDIs. HAL and Harbor are fitted separately. The reference bands $[2, 3]$ represent minimal detection limits used in laboratory measurement (CDER, 2024; Sheehan and Yost, 2026; Taleuzzaman, 2018), and are descriptive rather than calibrated thresholds.

5.4 Pooling diverse benchmarks can make model rankings more reliable

Agentic leaderboards often consist of multiple benchmarks from which model agentic capability is to be inferred. If the goal is to rank models on their ability to perform on diverse agentic tasks, pooling benchmarks that test different kinds of tasks provides more information about which performance differences persist across settings. At a fixed task budget, benchmark breadth averages condition-specific model advantages that within-benchmark replication leaves untouched (Corollary 3.2).

Table 3: **Agreement with external agent benchmarks.** Kendall's τ for the unweighted HAL mean and reliability-adjusted latent ($\widehat{\theta}$) model effect (median posterior) with $[95\%]$ CIs.

External Benchmark	Mean score	$\widehat{\theta}$	Difference	LLM Observations
BFCL v4	$0.429\,[-0.282,\,0.836]$	$\mathbf{0.714}\,[0.147,\,0.928]$	$+0.285$	7
Terminal-Bench 2.0	$0.524\,[-0.165,\,0.869]$	$\mathbf{0.810}\,[0.361,\,0.954]$	$+0.286$	7
SWE-bench Verified	$0.565\,[0.306,\,0.746]$	$\mathbf{0.765}\,[0.595,\,0.870]$	$+0.200$	20
τ^2-bench Core	$0.333\,[-0.515,\,0.852]$	$\mathbf{0.867}\,[0.383,\,0.977]$	$+0.534$	6
Mean τ	0.463	$\mathbf{0.789}$	$+0.326$	—

Pooling benchmarks can achieve more reliable model rankings at lower cost. At the same task budget, distributing evaluations across the nine-benchmark battery raises projected model-ranking reliability from approximately (0.44) for a single benchmark to (0.75) (Figure 3). A complementary pooled analysis of the Harbor Index (Shi et al., 2026) (Appendix A.2) supports the same qualitative pattern. Moreover, the full HAL battery costs more than ($47,000), while a balanced allocation achieves comparable reliability for approximately ($19,000). For an illustrative target of ($E\rho^2 \approx 0.714$), approximately (14) tasks per benchmark cost ($8,144), an estimated (83%) reduction. In Appendix Table 18, we corroborate the diminishing returns using task subsampling.

Task diversity leads to better model rank generalization. At the same task budget, projected model-ranking reliability increases from approximately 0.44 within a single benchmark to 0.75 across the nine-benchmark battery (Figure 3). The resulting shared model estimate also agrees more closely with all four held-out agent benchmarks than a simple average of benchmark scores (mean Kendall's τ: 0.789 vs. 0.463; Table 3; see § A.3 for data and contamination prevention). Together, these results suggest that pooling diverse benchmarks can help separate model differences that recur across settings from advantages specific to a particular benchmark or evaluation condition. These results provide convergent predictive evidence, not proof of a universal one-dimensional agent capability.[4]

Breadth is most useful when the design is connected with informative benchmarks. For these data, CORE-Bench Hard contains the greatest SNR (see Figure 7, § E.4), particularly with its own scaffold, CORE Agent (§ E.5). Removing this strongest discriminating signal increases pooled uncertainty more than any other LOO removals. Thus, having a benchmark with high model signal reliability can support anchoring a common scale through sufficient connectivity to other benchmarks.

Recommendation. When the intended claim concerns performance across different kinds of agentic tasks, pool benchmarks that represent that range rather than evaluating each benchmark exhaustively. Use the reliability analysis to determine how much evaluation is needed within each benchmark while preserving enough task diversity to support the broader cross-task claim. For imbalanced designs, prioritize identifying informative anchors to bridge conditions.

6 Conclusion

Agent evaluations should not be treated as having a single, intrinsic level of reliability. Reliability depends on what practitioners want to learn from them: the underlying model or a complete model–scaffold system, a ranking or an absolute score, and performance on one benchmark or across a broader range of agentic tasks. Our results show that these distinctions matter in practice and hence that evaluation design should follow the intended claim: Practitioners should first specify what they want a score or ranking to mean, then identify which sources of uncertainty prevent the evaluation from supporting that claim. More evaluation is useful only when it addresses those sources. Reliability analysis can therefore guide not only how evaluation results are interpreted, but also where additional evaluation effort is most valuable for differentiating a target model population.

[4]Internally, the pooled estimation strongly outperforms per-benchmark estimations in correlations of posterior rankings with benchmark-wise observed scores, [0.62,0.96] and [0.08,0.25] respectively (Table 9, § E.2). Yet posterior rank intervals remain wide; many neighboring models have ordering probabilities near $1/2$ (Figure 13), with some models moving across rank quartiles between reported and adjusted rankings (Figures 11 and 12).

REFERENCES

Anthropic. System Card: Claude Opus 4.7. System card, Anthropic, April 2026.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment, June 2025. URL https://arxiv.org/abs/2506.07982v1.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1), 2015. ISSN 1548-7660. doi: 10.18637/jss.v067.i01. URL http://www.jstatsoft.org/v67/i01/.

Aidan D. Bindoff. Partial pooling predicts cross-validation reliability: a closed-form triage and Rao-Blackwellised cure for hierarchical LOO, July 2026. URL http://arxiv.org/abs/2607.18836. arXiv:2607.18836 [stat.ME].

Robert L. Brennan. *Generalizability Theory*. Springer, New York, NY, 2001. ISBN 978-1-4419-2938-9 978-1-4757-3456-0. doi: 10.1007/978-1-4757-3456-0. URL http://link.springer.com/10.1007/978-1-4757-3456-0.

Paul-Christian Bürkner. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, 100:1–54, November 2021. ISSN 1548-7660. doi: 10.18637/jss.v100.i05. URL https://doi.org/10.18637/jss.v100.i05.

CDER. Q2(R2) Validation of Analytical Procedures, March 2024. URL https://www.regulations.gov/document/FDA-2022-D-1503-0002.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. ISSN 1939-1455. doi: 10.1037/h0040957.

Lee J. Cronbach and Richard J. Shavelson. My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, 64(3):391–418, June 2004. ISSN 0013-1644. doi: 10.1177/0013164404266386. URL https://doi.org/10.1177/0013164404266386.

Lee J. Cronbach, Goldine Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of behavioral measurements: theory of generalizability for scores and profiles*. Wiley, New York, 1972. ISBN 978-0-471-18850-6.

Jean-Paul Fox and Cees A. W. Glas. Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, 66(2):271–288, June 2001. ISSN 1860-0980. doi: 10.1007/BF02294839. URL https://doi.org/10.1007/BF02294839.

Michael Hardy and Yunsung Kim. Knowledge without Wisdom: Measuring Misalignment between LLMs and Intended Impact, April 2026. URL http://arxiv.org/abs/2603.00883. arXiv:2603.00883 [cs].

Michael Hardy, Anka Reuel, Lijin Zhang, Jodi M. Casabianca, Sang Truong, Yash Dave, Hansol Lee, Benjamin Domingue, and Sanmi Koyejo. AI Cartography: Mapping the Latent Landscape of AI Benchmark Ecosystems, May 2026. URL http://arxiv.org/abs/2605.25272. arXiv:2605.25272 [cs.AI].

Xavier A. Harrison. A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ*, 3:e1114, July 2015. ISSN 2167-8359. doi: 10.7717/peerj.1114. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4517959/.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6864–6890, 2024.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The twelfth international conference on learning representations*, 2023.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI Agents That Matter, July 2024. URL https://arxiv.org/abs/2407.01502v1.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S. Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, and Arvind Narayanan. Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation, October 2025. URL http://arxiv.org/abs/2510.11977. arXiv:2510.11977 [cs].

Pankaj Kumar and Subhankar Mishra. Robustness in large language models: A survey of mitigation strategies and evaluation metrics. *arXiv preprint arXiv:2505.18658*, 2025.

John Linacre and Benjamin Wright. Understanding Rasch measurement: Construction of measures from many-facet data. *Journal of applied measurement*, 3:486–512, February 2002.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying Variance in Evaluation Benchmarks, June 2024. URL http://arxiv.org/abs/2406.10229. arXiv:2406.10229.

Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10, December 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.02767. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.02767/full.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces, 2026. URL https://arxiv.org/abs/2601.11868. Version Number: 1.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

OpenAI. GPT-5.5 System Card. System card, OpenAI, April 2026.

Topi Paananen, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):16, February 2021. ISSN 1573-1375. doi: 10.1007/s11222-020-09982-2. URL https://doi.org/10.1007/s11222-020-09982-2.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 48371–48392. PMLR, October 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a Science of AI Agent Reliability, February 2026. URL http://arxiv.org/abs/2602.16666. arXiv:2602.16666 [cs].

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. *ArXiv*, abs/2502.06065, 2025. URL https://api.semanticscholar.org/CorpusID:276249536.

Maria L. Rizzo and Gábor J. Székely. DISCO analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2), June 2010. ISSN 1932-6157. doi: 10.1214/09-AOAS245. URL http://arxiv.org/abs/1011.2288. arXiv:1011.2288 [stat].

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation, June 2025. URL http://arxiv.org/abs/2505.10573. arXiv:2505.10573 [cs].

Richard J. Shavelson, Noreen M. Webb, and Glenn L. Rowley. Generalizability theory. *American Psychologist*, 44(6):922–932, 1989. ISSN 1935-990X. doi: 10.1037/0003-066X.44.6.922.

Terry L. Sheehan and Richard A. Yost. What's the Most Meaningful Standard for Mass Spectrometry: Instrument Detection Limit or Signal-to-Noise Ratio? | Spectroscopy Online, August 2026. URL https://www.spectroscopyonline.com/view/what-s-most-meaningful-standard-mass-spectrometry-instrument-detection-limit-or-signal-noise-ratio-0.

Lin Shi, Haowei Lin, Zixuan Zhu, Xiaoyue Zhou, Xiang Li, Xiangning Lin, Yaxuan Deng, Han Xu, Yuangang Li, Shanda Li, Zizhao Chen, Hanwen Xing, Harsh Raj, Bo Chen, Quan Shi, Steven Dillmann, Yipeng Gao, Puneesh Khanna, Ruofan Lu, Chao Beyond Zhou, Michael Yang, Robert Zhang, Siyuan Chai, Jiayu Chang, Yizhao Chen, Xiaokun Chen, Yiwei Dai, Wenting Yang, Hange Liu, Minghao Liu, Zihan Wang, Adnan El Assadi, Benedikt Stroebl, E. Kelly Buchanan, Han Meng, Junwei He, Longxuan Yu, Radin Shayanfar, Yukyung Lee, Zhikang Dong, Allen G Hart, Anjiang Wei, Anurag Kashyap, Arpandeep Khatua, Audrey Jixin Zheng, Chengrui Ma, David Heineman, Dubing Chen, Hai-Anh Trinh, Haishuo Fang, Hefan Zhang, Hui Shen, Issa Sugiura, Jiankai Sun, Jiechao Gao, Junhong Lin, Junnan Li, Kai Yang, Lei Hsiung, Maoyu Wang, Mengze Tang, Nabil Omi, Negin Raoof, Nicholas Edwards, Octavia Guo, Orfeas Menis Mastromichalakis, Pengliang Ji, Przemysław Hejman, Qi Qi, Qunshu Lin, Richard Zhuang, Rui Yang, Ruichen Zheng, Ryan Marten, Shaghayegh Fazliani, Shizheng Hou, Sicong Jiang, Sijie Li, Boqin Yuan, Michael Glass, Song Bian, Terry Yue Zhuo, Tianqing Wu, Tom Tang, Wanjia Zhao, Weihao Xuan, Wenhua Liang, Xian Liu, Xin Lan, Xuan Zhang, Xuandong Zhao, Yanchuan Tang, Yifan Jiang, Yijiang Li, Yitong Guan, Yizhi Li, Yonghui Liu, Yuheng Tang, Yujun, Mao, Yunfei Zhao, Yuxin Wang, Yuxuan Tang, Zhenheng Tang, Zhifei Li, Ziruo Wang, Ziyu She, Kaiyuan Liu, Iheb Chaabane, Yuxin Tang, Xiangyi Li, Satya Sai Srinath Namburi GNVV, Xinyue Zheng, Andy Konwinski, Boxuan Li, Leon Liangyu Chen, Alex Dimakis, Nicholas Carlini, Soroush Vosoughi, Sanmi Koyejo, Di He, Etash Guha, Benjamin Feuer, Mike Merrill, Ludwig Schmidt, and Alex Shaw. Harbor Adapters and Harbor-Index: Infrastructure and a Curated Meta-Dataset for Large-Scale Agentic Evaluation, September 2026. URL https://arxiv.org/abs/2609.04298. Version Number: 3.

Quan Shi, Michael Tang, Karthik Narasimhan, and Shunyu Yao. Can language models solve olympiad programming? *arXiv preprint arXiv:2404.10952*, 2024.

Zachary S Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. Corebench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel Dsouza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The Leaderboard Illusion. *Advances in Neural Information Processing Systems*, 38, April 2026. doi: 10.52202/085713-2620. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/70a93f260a51123b3c0e33ecd1b4de97-Abstract-Datasets_and_Benchmarks_Track.html.

Gábor J. Székely and Maria L. Rizzo. The Energy of Data. *Annual Review of Statistics and Its Application*, 4(1):447–479, March 2017. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-060116-054026. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-060116-054026.

Mohamad Taleuzzaman. Limit of Blank (LOB), Limit of Detection (LOD), and Limit of Quantification (LOQ). *Organic & Medicinal Chemistry International Journal*, 7(5), August 2018. ISSN 24747610. doi: 10.19080/OMCIJ.2018.07.555722. URL https://juniperpublishers.com/omcij/OMCIJ.MS.ID.555722.php.

Minyang Tian, Luyu Gao, Shizhuo D Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists. *Advances in Neural Information Processing Systems*, 37:30624–30650, 2024.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, September 2017. ISSN 1573-1375. doi: 10.1007/s11222-016-9696-4. URL https://doi.org/10.1007/s11222-016-9696-4.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto Smoothed Importance Sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024. ISSN 1533-7928. URL http://jmlr.org/papers/v25/19-556.html.

Ruipeng Wang, Yuxin Chen, Yukai Wang, Chang Wu, Junfeng Fang, Xiaodong Cai, Qi Gu, Hui Su, An Zhang, Xiang Wang, et al. Agentnoisebench: Benchmarking robustness of tool-using llm agents under noisy condition. *arXiv preprint arXiv:2602.11348*, 2026.

Wen-Chung Wang and Mark Wilson. Exploring Local Item Dependence Using a Random-Effects Facet Model. *Applied Psychological Measurement*, 29(4):296–318, July 2005. ISSN 0146-6216. doi: 10.1177/0146621605276281. URL https://doi.org/10.1177/0146621605276281.

Tianci Xue, Weijian Qi, Tianneng Shi, Chan Hee Song, Boyu Gou, Dawn Song, Huan Sun, and Yu Su. An Illusion of Progress? Assessing the Current State of Web Agents, October 2025. URL http://arxiv.org/abs/2504.01382. arXiv:2504.01382 [cs.AI].

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8938–8968, 2024.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

APPENDIX

LIMITATIONS AND FUTURE WORK

Reliability is necessary but not sufficient for valid evaluation. The bounds we estimate describe how consistently a benchmark differentiates the objects it ranks, not whether the rankings correspond to underlying capability or to any external ground truth: a benchmark can be reliable but invalid. The object-of-measurement problem we document is in this sense a construct validity problem in disguise, since reliability against an undefined construct cannot be assessed in principle (Cronbach and Meehl, 1955; Salaudeen et al., 2025). Future work should complement the reliability framework with validity studies, including external-criterion validation and cross-benchmark transfer analyses.

Scope conditions on the headline ceiling. The reliability bounds we report are joint properties of the benchmark, the scaffold sample, the task sample, and the competitor set being ranked. A clustered frontier-model pool depresses σ^2_M regardless of benchmark quality, lowering reliability not because the benchmark is poorly designed but because it is being used to distinguish systems too close for its resolution; hence, the ceiling we find should be read as a statement about current leaderboards for the current frontier-model competitor set. Scaffolds are selected artifacts, often co-developed with the benchmarks they evaluate, and per-benchmark counts ($n_a \in \{2, 3\}$) reflect standard practice across the field rather than a feature of HAL; pooling across benchmarks in Equation 6 partially addresses but per-benchmark scaffold-variance claims remain weaker, and pooling itself depends on HAL-GENERALIST as the cross-benchmark harness anchor (Appendix F). Future work should construct evaluation datasets that systematically span larger scaffold libraries, and should track how the ceiling shifts as the model pool evolves.

Scope of Agent Scaffolds and Systems. These results are conditional on the sampled model pool and scaffold set. A compressed frontier depresses σ^2_M and thus reliability independent of benchmark quality; a richer scaffold sample would sharpen the model–scaffold decomposition. Our claims are therefore about *these leaderboards for this competitor set*, and generalize to new models and scaffolds only insofar as the sampled conditions represent the intended universe. We view stating this scope explicitly as part of reliable practice rather than a caveat to it. Inference remains conditional on the represented evaluation population; partial pooling does not by itself correct selective evaluation or reporting.

Latent-scale reliability. We estimate reliability on the latent logit scale, while leaderboards report observed proportions. The two scales do not translate one-to-one, so our numbers should be read as approximate bounds on observed-scale rank stability. Future work could entail a sensitivity analysis comparing the scales.

Raising reliability bounds. The empirical scope of this work is the characterization of reliability bounds rather than their displacement. Validating interventions that raise the ceiling is the subject of subsequent work, since each candidate intervention introduces its own measurement-design tradeoffs. Candidate interventions include task selection using item discrimination methods, scaffold sampling under a defined scaffold family, and task-quality auditing; empirical evaluation of these, alongside methods that reduce the measurement cost of the recommendations, are the most direct extensions.

A DATASET DESCRIPTIONS

A.1 HAL DATASET

The Holistic Agent Leaderboard (HAL) dataset[5] is a standardized, cost-aware, and third-party evaluation platform and dataset initiative developed by the SAgE (Science of Agent Evaluation) research group at Princeton University. The formatted and cleaned tasks-level of this dataset is publicly released [6].

[5] https://hal.cs.princeton.edu/

[6] https://huggingface.co/datasets/razam2/hal-response-matrix

A.1.1 BENCHMARKS

AssistantBench Tasks consist of time-consuming, busy-work tasks that an average person may face, seeking information from the web.

> *Which gyms near Tompkins Square Park (<200m) have fitness classes before 7am?*

CORE-Bench Hard Tasks ask an agent to computationally reproduce specific quantitative results from a published scientific paper, given a code repository, dataset, and research paper.

> *Run the main.py file three times. First, with `config/uci.json`, the preprocessing task, and the CTGCN-C method. Second, with `config/uci.json`, the embedding task, and the CTGCN-C method. Third, using Python3 with `config/uci.json` and the link-pred task.*

GAIA GAIA consists of tasks requiring multi-step tool use — including web browsing, code execution, file reading (PDFs, spreadsheets, audio), and multimodal understanding.

> *(Given an Excel file) The attached Excel file contains the sales of menu items for a local fast-food chain. What were the total sales that the chain made from food (not including drinks)? Express your answer in USD with two decimal places.*

OnlineMind2Web Online Mind2Web is the live, online version of Mind2Web. It does not rely on cached pages and allows for real-time testing against dynamic, evolving web interfaces. All tasks are sourced from 136 popular websites to reflect authentic user workflows.

> *Browse used Audi cars made before 2015 and sort by lowest price on KBB. Website: `https://www.kbb.com/`*

SciCode Tasks consist of research-level coding problems decomposed into subproblems drawn from actual scientific work across physics, chemistry, biology, math, and materials science.

> *Main problem: Reproduce the Chern number phase diagram of the Haldane model on a hexagonal lattice. Subproblem 1.1: Write a Haldane model Hamiltonian on a hexagonal lattice, given: wavevector components k_x and k_y, lattice spacing a, nearest-neighbor coupling constant t_1, next-nearest-neighbor coupling constant t_2, phase φ for next-nearest-neighbor hopping, and on-site energy m. Output: a 2×2 matrix. Subproblem 1.2: Calculate the Chern number using the Haldane Hamiltonian, given the grid size δ for discretizing the Brillouin zone...*

ScienceAgentBench ScienceAgentBench is a benchmark for evaluating the ability of language agents to conduct data-driven scientific discovery.

> *(Given separate training and test datasets) Train a multitask model on the Clintox dataset to predict a drug's toxicity and FDA approval status. Save the test set predictions, including the SMILES representation of drugs and the probability of positive labels, to "`pred_results/clintox_test_pred.csv`".*

SWE-bench Verified Mini A subset of SWE-bench Verified where each task gives the agent a real GitHub issue description and a full Python repository, and requires the agent to produce a code patch that makes failing unit tests pass.

> *(Given a repo and issue description) Navigate the repository, identify the relevant code path in `django/db/models/query.py`, and produce a `.patch` file that makes the FAIL_TO_PASS unit tests pass without breaking existing PASS_TO_PASS tests.*

τ-bench Airline Tasks simulate a realistic airline customer service scenario in which a human user (played by another LLM) contacts an agent with requests like rebooking a flight, adding a passenger,

or canceling a reservation. The agent must follow a detailed airline policy document and a limited set of callable functions.

> *(Given the airline's policy and pre-defined functions) Your user id is* `daiki_muller_1116`*. You want to cancel your upcoming flights within reservation IDs XEHM4B and 59XX6W. If the agent says either reservation has basic economy flights, ask to upgrade them to economy first and then cancel them. You are very persistent and terse but clear. After the third agent message, also ask whether you have any other upcoming flights and what their total cost is.*

USACO Tasks are problems from the USA Computing Olympiad, spanning four difficulty tiers (Bronze through Platinum).

> *Farmer John has N cows ($2 \leq N \leq 10^5$), each liking exactly one type of hay h_i. He can host focus groups over contiguous ranges of cows — if more than half the cows in a group prefer the same type, all cows switch to that type. He wants to know which types of hay can become universally liked. Given T test cases, each with N and a list of h_i values, output all achievable universal hay types in increasing order, or -1.*

A.1.2 SCAFFOLDS

The scaffolds included in the HAL benchmark dataset are in Table 6. For each benchmark, the scaffold consist of a contrast between a generalist scaffold (e.g., HAL generalist, Claude Code) and and a specialist scaffold tailored to the specific benchmark (e.g., SWE agent, τ-bench tool-calling).

A.2 HARBOR INDEX DATASET FOR DECISION STUDY CORROBORATION

Harbor-Index[7] is a curated meta-dataset and benchmark containing 82 difficult and diverse tasks designed for evaluating AI language model agents. Distilled from a pool of over 6,000 candidate tasks across 54 benchmarks, the final dataset features 82 high-quality tasks spanning 29 benchmarks and seven domains (including software engineering, scientific research, tool use, mathematics, data analytics, and security). It was built by passing candidate tasks through difficulty filtering, automated AI audits, human reviews, and iterative audit-and-fix loops to weed out structurally broken or flawed tasks.

The final dataset lacks sufficient per-benchmark task representation that prevent inclusion in all the analyses of this paper. Descriptive statistics are in Table 4. Thus, to estimate both task and benchmark effects, we take the subset of benchmarks with at least three items. The cross-benchmark model of Equation 6 is fit and produces the sample the posterior draws used in Figure 3b. For our analyses, we use their verified and judged outcome classifications as the measure of task success (i.e., if the verifier marked a reward as a false negative and the judge determined it was a "true solve" it was marked as successful). Because the dataset suffers from a small quantity of total tasks, both the proportion of variance due to persistent differences in LLM and overall reliability begins to drop more sharply if we take the subset of benchmarks with at least four items, decreasing further with the subset that has five items. We conjecture this pattern may be in part due to the nonrandom nature of the item selection process used to represent each benchmark.

A.3 EXTERNAL VALIDATION DATASETS FOR MODEL RANKINGS

We compare the posterior-median model effect $\widehat{\theta}_m$ with four contemporaneous external benchmarks containing at least six overlapping LLMs and which report individual task-level performance by model. To prevent data leakage for the external SWE-bench and τ^2-bench, the semantically corresponding HAL benchmark is excluded before estimating both $\widehat{\theta}_m$ and the raw-score baseline (see § E.4), as described in the details of the data and collection below.

- **The Berkeley Function Calling Leaderboard** (BFCL) V4 (Patil et al., 2025) is a benchmark designed to evaluate the tool use, API execution, and agentic capabilities of large language models.[8]. We use BFCL's holistic score for function-calling models.

[7] `https://harbor-index.org/`

[8] `https://gorilla.cs.berkeley.edu/leaderboard.html`

Table 4: Descriptive statistics from the complementary Harbor Index dataset. **Bolded** datasets have at least three tasks and were used in the analysis of benchmark and item effects on agent system signal.

Benchmark	Models	Tasks	Scaffold	Model-Scaffold Pair	Mean Score
algotune	9	5	4	18	0.078
arcagi2	9	5	4	18	0.067
bigcodebench	9	1	4	18	0.222
bixbench	9	5	4	18	0.122
codepde	9	1	4	18	0.000
cybergym	9	2	4	18	0.1 39
dacode	9	1	4	18	0.111
featurebench	9	4	4	18	0.042
gaia	9	3	4	18	0.185
gaia2	9	5	4	18	0.051
gpqadiamond	9	1	4	18	0.056
gso	9	7	4	18	0.024
hle	9	8	4	18	0.079
labbench	9	4	4	18	0.182
omnimath	9	2	4	18	0.278
qcircuitbench	9	1	4	18	0.000
replicationbench	9	1	4	18	0.444
scicode	9	3	4	18	0.019
skillsbench	9	2	4	18	0.167
sldbench	9	1	4	18	0.056
spider2	9	2	4	18	0.000
swebenchpro	9	4	4	18	0.139
swebenchverified	9	5	4	18	0.067
swelancer	9	2	4	18	0.139
swesmith	9	1	4	18	0.111
swtbenchverified	9	1	4	18	0.333
tb	9	3	4	18	0.241
usaco	9	1	4	18	0.000
widesearch	9	1	4	18	0.000

- **Terminal-Bench 2.0** (Merrill et al., 2026) is an evaluation suite and benchmark designed to test how well AI agents perform complex, multi-step tasks inside a real command-line interface (CLI).[9]. To compare LLMs, we average each model's score across reported agent runs.
- **SWE-bench Verified** (Jimenez et al., 2023) is the superset of tasks that are associated with the HAL dataset SWE-bench Verified Mini. The data of several third-party external leaderboard providers[10] were collected. Accuracy scores were averaged across leaderboards for overlapping models for the standard Kendall's τ, and where left separate for the multilevel partial correlation (Appendix A.3.1). When performing the analyses in Section **??**, the correlations were calculated after removing SWE-bench Verified Mini from the analysis. The description of this process based on the pooled model (Eq. 6) can be found in Appendix E.4.
- **τ^2-bench Core** (Barres et al., 2025) is an evaluation framework and benchmark for conversational AI and customer service agents that tests performance in a "dual-control" environment where both the AI agent and the user take active actions. It contains tasks that evaluate LLM agents on retail, airline (of which HAL's τ-bench Airline is a subset), and telecom

[9] https://hub.harborframework.com/datasets/terminal-bench/terminal-bench-2/latest?tab=leaderboard&leaderboard=2-0

[10] https://www.swebench.com/, https://www.vals.ai/benchmarks/swebench, https://llm-stats.com/benchmarks/swe-bench-verified

customer-service tasks, where the agent and the user both act on the world.[11] We use the overall score across all of these domains and, as with SWE-bench Verified, we only calculate correlations with HAL data after removing τ-bench Airline. The description of this process based on the pooled model (Eq. 6) can be found in Appendix E.4.

A.3.1 Combined External Data Multilevel Partial Correlation

In addition to traditional Kendall's τ_b correlations, a multilevel partial correlation was calculated to clarify the relationships with greater statistical power. For LLM m evaluated on benchmark b and external leaderboard providers ℓ, let $R_{mb\ell}$ denote its observed performance rank and $P_{mb\ell}$ the rank induced by the calculated proxy. Benchmark- and leaderboard-provider-level heterogeneity can be modeled through the rank-based mixed-effects specifications

$$R_{mb\ell} = \mathbf{x}_{mb\ell}^{\top}\boldsymbol{\beta}_R + u_b^R + v_\ell^R + \varepsilon_{mb\ell}^R, \qquad P_{mb\ell} = \mathbf{x}_{mb\ell}^{\top}\boldsymbol{\beta}_P + u_b^P + v_\ell^P + \varepsilon_{mb\ell}^P,$$

where $u_b^{(\cdot)}$ and $v_\ell^{(\cdot)}$ are random effects for benchmarks and leaderboards, respectively, and $\mathbf{x}_{mb\ell}$ contains any control variables. The inclusion of benchmark and leaderboard random effects removes the need to average across the SWE-bench Verified leaderboards, increasing the overall number of observations to 52 for this analysis. The multilevel partial Kendall correlation is then defined as the concordance between the adjusted residual ranks,

$$\tau_{\text{partial}} = \mathbb{E}\Big[\operatorname{sgn}\Big(\tilde{R}_a - \tilde{R}_{a'}\Big)\operatorname{sgn}\Big(\tilde{P}_a - \tilde{P}_{a'}\Big)\Big],$$

where $\tilde{R}_{mb\ell} = R_{mb\ell} - \mathbf{x}_{mb\ell}^{\top}\hat{\boldsymbol{\beta}}_R - \hat{u}_b^R - \hat{v}_\ell^R$ and $\tilde{P}_{mb\ell} = P_{mb\ell} - \mathbf{x}_{mb\ell}^{\top}\hat{\boldsymbol{\beta}}_P - \hat{u}_b^P - \hat{v}_\ell^P$. Thus, τ_{partial} measures agreement between the LLM rankings and the proxy rankings after accounting for observed covariates and clustering attributable to benchmarks and leaderboards. The results are in Table 5.

Table 5: **Agreement with external agent benchmarks.** Kendall's τ for the unweighted HAL mean and reliability-adjusted latent ($\widehat{\theta}$) model effect (median posterior) with $[95\%]$ CIs.

External Benchmark	Mean score	$\widehat{\theta}$	Difference	LLM Observations
Multilevel τ_{partial}	$0.430\,[0.266,\ 0.570]$	$\mathbf{0.632}\,[0.507,\ 0.732]$	$+0.202$	52

B Coverage Matrix

Below is the list of agents used to test each of the benchmarks. Each model was not uniformly tested with each benchmark and agent. As shown in Table 6, about nearly all (11/13) agents have only been tested on a single benchmark. Similarly, many models (12/54) are tested with a single agent. The full breakdown of which models, agents and benchmarks have been tested can be seen in Figure 5.

Table 6: Agent harness/scaffold by benchmark.

Benchmark	**Agent Names**
AssistantBench	hal_generalist, browser-use
CORE-Bench Hard	hal_generalist, coreagent, claude_code
GAIA	hal_generalist, hf_open_deep_research
OnlineMind2Web	browser-use, seeact
SciCode	hal_generalist, tool_calling_agent, scicode_zero
ScienceAgentBench	hal_generalist, sab_selfdebug
SWE-bench Verified Mini	hal_generalist, sweagent
τ-bench Airline	hal_generalist, taubench_tool_calling, taubench_fewshot
usaco	hal_generalist, usaco_episodic_semantic

[11] https://taubench.com/leaderboard?benchmark=core

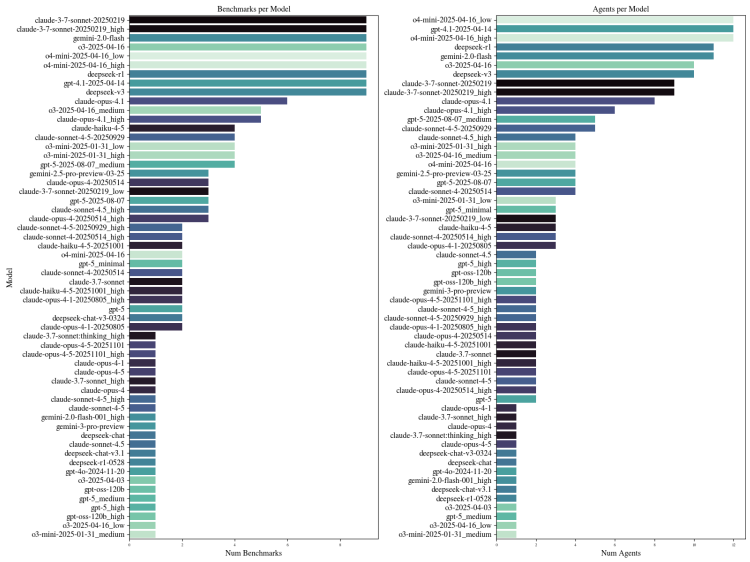


Figure 4: Number of benchmarks tested (left) and agents tested (right) per model.

C Methods, Estimation, and Computational Details

This appendix reports prior distributions, parameterization, sampler settings, convergence diagnostics, posterior predictive checks, and the treatment of repeated observations. It also gives reliability expressions for the observed unbalanced allocation. For each posterior draw, design-specific error is computed using the realized cell weights rather than the equal-allocation approximations used for the main D-studies.

We provide the full random-effects structure for each model using standard mixed-model notation. For the main body studies, we estimate all variance components jointly using Bayesian generalized linear mixed models, which provide partial pooling for sparsely observed cells and propagate uncertainty into the resulting generalizability coefficients.

C.1 Estimation Details

C.1.1 Leaderboard-level pooled model

Let $y_{bima} \in \{0, 1\}$ denote success on benchmark b, task i, model m, and scaffold a. We assume

$$y_{bima} \mid p_{bima} \sim \text{Bernoulli}(p_{bima}), \qquad \text{logit}(p_{bima}) = \eta_{bima},$$

with latent linear predictor (from Equation 6)

$$\eta_{bima} = \mu + u_b + u_{bi} + u_m + u_a + u_{bim} + u_{bia} + u_{ma} + u_{bm} + u_{ba} + u_{bma}.$$

Here bi identifies a task nested within its benchmark. Each term is a mean-zero random intercept,

$$u_{F,j} \mid \sigma_F \sim \mathcal{N}(0, \sigma_F^2), \qquad F \in \{b, bi, m, a, bim, bia, ma, bm, ba, bma\},$$

and random-effect families are conditionally independent. This decomposition separates persistent model and scaffold differences from variation attributable to benchmark choice, task composition, and model–scaffold compatibility. In particular, u_{bm} captures benchmark-dependent model performance, while u_{bma} captures benchmark-specific compatibility between models and scaffolds; both can alter rankings even when average model effects are unchanged.

The model was fitted to 29,923 observations spanning 9 benchmarks, 1,117 benchmark–task units, 54 models, and 13 scaffolds. Because the likelihood is Bernoulli-logit, all variance components are defined on the latent log-odds scale. When an observation-level residual is required for a generalizability coefficient, the conventional logistic variance $\pi^2/3$ is used. Reliability quantities are

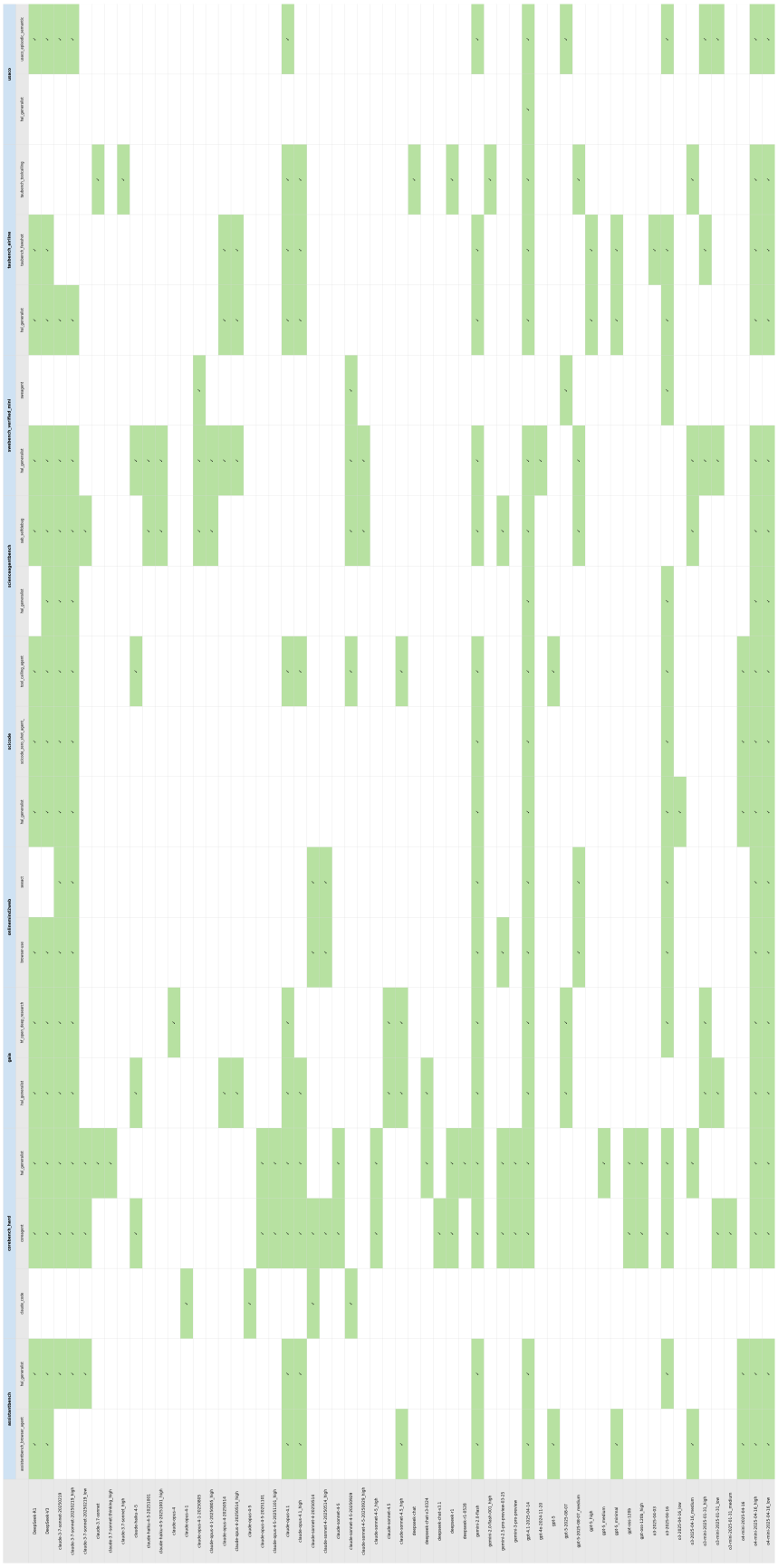


Figure 5: Matrix indicating all the variations of benchmarks, scaffold and models (with reasoning-effort) that were included in our dataset.

computed separately for every posterior draw, rather than from ratios of posterior mean variances, thereby preserving uncertainty in nonlinear variance decompositions.

`brms` specification:

```
brm(score ~ 1 + (1|benchmark) + (1|benchmark:task_id)
    + (1|model_name) + (1|agent_name)
    + (1|benchmark:task_id:model_name)
    + (1|benchmark:task_id:agent_name)
    + (1|model_name:agent_name)
    + (1|benchmark:model_name)
    + (1|benchmark:agent_name)
    + (1|benchmark:model_name:agent_name),
    family = bernoulli("logit"), ...)
```

C.1.2 BENCHMARK-LEVEL DECOMPOSITION MODELS

To quantify the signal supplied by each benchmark, we also fit a separate model within every benchmark using Equation 2:

$$y_{ima} \sim \text{Bernoulli}(p_{ima}),$$
$$\text{logit}(p_{ima}) = \mu + u_i + u_m + u_a + u_{im} + u_{ia} + u_{ma}.$$

These fits distinguish stable model variation from task-, scaffold-, and interaction-driven variation within an instrument. Thus, a benchmark with many observations need not be highly informative for ranking models: its effective signal depends on the posterior magnitude of model-related variance relative to the variance induced by tasks, scaffolds, and their interactions.

`brms` specification:

```
brm(score ~ 1 + (1|task_id) + (1|model_name) + (1|agent_name)
    + (1|task_id:model_name) + (1|task_id:agent_name)
    + (1|model_name:agent_name),
    family = bernoulli("logit"), ...)
```

C.1.3 PRIORS AND POSTERIOR COMPUTATION

We used the default weakly informative `brms` priors. For every group-level standard deviation,

$$\sigma_F \sim \text{Student-}t^+(3, 0, 2.5),$$

where t^+ denotes truncation to $\sigma_F \geq 0$. Equivalently, the estimated variance component is σ_F^2. The prior is concentrated near modest latent-scale heterogeneity while retaining sufficiently heavy tails for large benchmark or task effects. The intercept used the corresponding weakly informative $\text{Student-}t(3, 0, 2.5)$ prior. These priors regularize components supported by few levels—notably benchmark and scaffold effects—without forcing them toward equality.

Posterior sampling used Stan's Hamiltonian Monte Carlo implementation through `brms` (Bürkner, 2021), with the conservative target average proposal acceptance probability during the adaptation period for the sampler `adapt_delta` $= 0.95$ without any resultant divergent transitions. The full model used six chains of $9{,}000$ iterations, including $4{,}000$ warm-up iterations, with thinning by eight, yielding $3{,}750$ retained draws. Each benchmark-specific model used four chains of $2{,}000$ iterations, including $1{,}000$ warm-up iterations, with thinning by three. On an Apple M1 Max, the full fit required approximately 5.5 hours, while a benchmark-specific fit required approximately 6.5 minutes on average.

For the full model, all reported split-$\widehat{R}$ values rounded to 1.00; bulk effective sample sizes for variance components ranged from $1{,}559$ to $3{,}743$, and tail effective sample sizes ranged from $2{,}062$ to $3{,}783$. Benchmark-specific fits were also inspected individually. For example, the largest split-$\widehat{R}$ in the USACO fit was 1.03, with uncertainty retained in all downstream posterior summaries rather than suppressed through plug-in estimates.

Table 7: Posterior summary for the full leaderboard-level model (Eq. 6). Group-level entries are standard deviations of random intercepts on the latent log-odds scale. Intervals are equal-tailed 95% posterior credible intervals.

Effect	Levels	Estimate	SD	95% CrI	$\widehat{R}$	Bulk ESS	Tail ESS
Group-level standard deviations							
Scaffold, σ_a	13	0.79	0.44	$[0.06,\ 1.74]$	1.00	3454	3360
Benchmark, σ_b	9	2.45	0.73	$[1.39,\ 4.21]$	1.00	3743	3783
Benchmark × scaffold, σ_{ba}	21	0.69	0.36	$[0.06,\ 1.43]$	1.00	3580	3615
Benchmark × model, σ_{bm}	182	0.52	0.17	$[0.12,\ 0.80]$	1.00	1771	2110
Benchmark × model × scaffold, σ_{bma}	287	0.55	0.14	$[0.24,\ 0.79]$	1.00	1559	2062
Task within benchmark, σ_{bt}	1117	2.41	0.09	$[2.25,\ 2.58]$	1.00	3309	3350
Task × scaffold in benchmark, σ_{bta}	2394	1.03	0.05	$[0.93,\ 1.14]$	1.00	3445	3540
Task × model in benchmark, σ_{btm}	18856	0.88	0.07	$[0.75,\ 1.01]$	1.00	2513	3241
Model, σ_m	54	0.78	0.14	$[0.52,\ 1.07]$	1.00	3152	3550
Model × scaffold, σ_{ma}	209	0.36	0.14	$[0.05,\ 0.61]$	1.00	1852	2376
Population-level coefficient							
Intercept, μ	—	-1.95	0.88	$[-3.62, -0.17]$	1.00	3693	3574

Note. "Levels" denotes the number of observed levels of each grouping factor. Estimate and SD are the posterior mean and posterior standard deviation, respectively. For group-level effects, the estimate summarizes the random-intercept standard deviation σ; for the intercept, it summarizes the coefficient itself. ESS denotes effective sample size. "Scaffold" corresponds to `agent_name` in the estimation data.

Convergence. All parameters across all models fit and all estimates derived from posterior draws throughout this study achieve $\hat{R} < 1.05$, the standard threshold for adequate mixing. For the 95% of estimates $\widehat{R}$ values rounded to 1.00. For the leaderboard-level model, complete convergence information is in Table 7

C.1.4 Leave-one-benchmark-out stability

To assess whether the leaderboard-level decomposition was dominated by any single instrument, we refit Equation 6 after removing each benchmark in turn. These are full Bayesian refits, not importance-sampling approximations. Each refit used four chains of $4{,}000$ iterations, including $2{,}000$ warm-up iterations, with thinning by five, yielding $1{,}600$ retained draws. Stability was evaluated by comparing the posterior distributions of variance components and derived reliability quantities with those from the complete-data fit. This analysis directly tests whether conclusions about model, scaffold, and benchmark contributions persist under changes to the benchmark ecosystem.

C.2 Bayesian Estimation for Linear Mixed Effect Models

To establish methodological sensitivity, Equation 6 was estimated using a Bayesian linear mixed effect model on the observation scale. Additional detail about the differences in methods and results can be found in Appendix H. All linear estimations of this use the same hyperparameters as above, except that the family is Gaussian with an identity link.

C.3 Frequentist Linear Mixed Effects (LME)

We fit a Gaussian identity-link model via restricted maximum likelihood (REML). Although the binary outcome violates normality, LME provides a familiar baseline and is the most commonly used variance-component estimator in G-theory applications. Variance components are extracted directly from the REML fit.

C.4 Frequentist Generalized Linear Mixed Effects (GLME)

A logistic mixed-effects model with a Bernoulli likelihood and logit link is fit via Laplace approximation to the marginal likelihood, with parameter estimation by penalized iteratively reweighted least squares (PIRLS). This respects the binary nature of the data but estimates are on the logit scale; we

Table 8: Reliability estimands. Each row applies $E\rho_o^2 = \sigma_o^2/(\sigma_o^2 + \sigma_\delta^2)$, $\text{SNR}_o = \sigma_o^2/\sigma_\delta^2$, but changes the object whose ordering should generalize. The n_f terms denote equal allocation over facet f.

Estimand	Object of inference	σ_o^2	$\sigma_\delta^2(\mathcal{D})$	Eq.
$E\rho^2_{M(b)}$, $\text{SNR}_{M(b)}$	Does benchmark b preserve model order?	σ^2_M	$\frac{\sigma^2_{IM}}{n_i} + \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{IMA,e}}{n_i n_a}$	(3)
$E\rho^2_{MA(b)}$	Does benchmark b preserve system order?	$\sigma^2_M + \sigma^2_A + \sigma^2_{MA}$	$\frac{\sigma^2_{IM} + \sigma^2_{IA} + \sigma^2_{IMA,e}}{n_i}$	(4)
$\rho^{(b)}_{AA'}$	Do scaffolds induce the same model order?	$\sigma^2_M + \frac{\sigma^2_{IM}}{n_i}$	$\sigma^2_{MA} + \frac{\sigma^2_{IMA,e}}{n_i}$	(5)
$E\rho^2_M$, SNR_M	Does the leaderboard preserve model order?	σ^2_M	$\frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{BMA}}{n_b n_a} + \frac{\sigma^2_{IM[B]}}{n_b n_i} + \frac{\sigma^2_{BIMA,e}}{n_b n_i n_a}$	(7)
$\lim_{n_i\to\infty} E\rho^2_{M(\cdot)}$	Does having unlimited tasks preserve model order?	σ^2_M	$\frac{\sigma^2_{MA}}{n_a}$ or $\frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{BMA}}{n_b n_a}$	(8)
$\Delta\rho^2_{(M-A)(b)}$	Do models have more benchmark-relevant signal than scaffolds?	$\sigma^2_M - \sigma^2_A$	$\sigma^2_{MA} + \frac{\sigma^2_{IM} + \sigma^2_{IA} + \sigma^2_{IMA,e}}{n_i}$	(11)
$\Delta\rho^2_{I(M-A)(b)}$	Do models have more task-level signal than scaffolds?	$\sigma^2_M + \frac{\sigma^2_{IM}}{n_i} - \sigma^2_A - \frac{\sigma^2_{IA}}{n_i}$	$\sigma^2_{MA} + \frac{\sigma^2_{IM} + \sigma^2_{IA} + \sigma^2_{IMA,e}}{n_i}$	(12)
$\Delta\rho^2_{M-A}$	Do models have more leaderboard-relevant signal than scaffolds?	$\sigma^2_M - \sigma^2_A$	$\sigma^2_{MA} + \frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{BA}}{n_b} + \frac{\sigma^2_{BMA}}{n_b} + \frac{\sigma^2_{IM[B]}}{n_b n_i} + \frac{\sigma^2_{IA[B]}}{n_b n_i} + \frac{\sigma^2_{BIMA,e}}{n_b n_i}$	(13)
$\Delta\rho^2_{B(M-A)}$	Do models have more benchmark-level signal than scaffolds?	$\sigma^2_M + \frac{\sigma^2_{BM}}{n_b} - \sigma^2_A - \frac{\sigma^2_{BA}}{n_b}$	$\sigma^2_{MA} + \frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{BA}}{n_b} + \frac{\sigma^2_{BMA}}{n_b} + \frac{\sigma^2_{IM[B]}}{n_b n_i} + \frac{\sigma^2_{IA[B]}}{n_b n_i} + \frac{\sigma^2_{BIMA,e}}{n_b n_i}$	(14)
$\Delta\rho^2_{I[B](M-A)}$	Do models have more task-level signal than scaffolds?	$\sigma^2_M + \frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{IM}}{n_b n_i} - \sigma^2_A - \frac{\sigma^2_{BA}}{n_b} - \frac{\sigma^2_{IA}}{n_b n_i}$	$\sigma^2_{MA} + \frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{BA}}{n_b} + \frac{\sigma^2_{BMA}}{n_b} + \frac{\sigma^2_{IM[B]}}{n_b n_i} + \frac{\sigma^2_{IA[B]}}{n_b n_i} + \frac{\sigma^2_{BIMA,e}}{n_b n_i}$	(15)

convert variance proportions by computing the share of total variance (including the logistic residual variance $\pi^2/3$) attributable to each component. Both the Linear and Generalized Linear models were estimated using `lme4` (Bates et al., 2015).

C.5 Nonparametric Distance Components (DISCO)

We estimate a nonparametric variance decomposition (see Appendix H.2) using distance components (Rizzo and Székely, 2010; Székely and Rizzo, 2017) from the energy-statistics literature. DISCO decomposes total dispersion—measured by pairwise Euclidean distances—into between- and within-group components without distributional assumptions. For a single facet with K groups,

$$\mathcal{S}_{\text{total}} \;=\; \mathcal{S}_{\text{between}} + \mathcal{S}_{\text{within}}, \tag{10}$$

where $\mathcal{S}$ denotes the energy-based dispersion statistic. The G-coefficient analog is $E\rho^2_{\text{DISCO}} = \mathcal{S}_p/(\mathcal{S}_p + \mathcal{S}_{\text{within}})$. DISCO captures nonlinear relationships and is robust to the heavy skewness observed in the Bayesian posteriors, particularly for facets with few levels (e.g., agents). However, it does not guarantee a positive $\sigma^2_\epsilon/\mathcal{S}_{\text{global within}}$ term, so its reported "variance"/dispersion shares sum to one across the total explained dispersion.

C.6 Details On Compute-Usage

Estimations of the main variance decomposition models used in the body of the paper took 8 total hours on an Apple M1 Max. LOO ablation studies in Appendix E took 57 hours. Methodological contrasts reported in Appendix H took 2 hours for frequentist estimations (both linear mixed effect models and generalized mixed effect models), Bayesian linear mixed effect models took 8 hours, and nonparametric distance components estimates took 18 hours.

C.7 Posterior Estimands

Table 8 contains the full set of estimands found in the paper.

D AI Use

AI was used during the writing phase of this study. GPT 5.6 Sol was used after initial drafts to reduce the text length of several sections in the main body, all of which required further editing after use. Drafts of several sections of Appendix F were created from research notes using GPT 5.5, which was revised and then rewritten using combinations of text from GPT 5.5, GPT 5.6 Sol, and Claude Opus 4.8. Paragraphs describing tabular results in Appendices E.4–E.6 were revised using GPT 5.5. Finally, the 100% human-written code used in this study was refactored with comments added to support reproducibility using Claude Opus 4.8.

E Ablations and Sensitivity Analyses

The benchmark ecosystem is sparse, unbalanced, and only partially connected. Tasks are nested within benchmarks, most model–scaffold combinations are absent, and many interaction levels are observed only once. Consequently, a fully crossed variance decomposition could in principle be driven by a small number of influential observations or by a particularly informative benchmark. We therefore examine robustness at four complementary levels: pointwise predictive stability, stability of latent rankings, leave-one-benchmark-out sensitivity, and leave-one-benchmark–scaffold-out sensitivity.

E.1 Reference indexed variance decomposition

For observation n, let $b[n]$, $i[n]$, $m[n]$, and $a[n]$ denote its benchmark, task, model, and scaffold. The reference model with additional indexing from equation 6 is

$$\begin{aligned} Y_n &\sim \text{Bernoulli}(p_n), \\ \text{logit}(p_n) &= \mu + u^B_{b[n]} + u^I_{b[n],i[n]} + u^M_{m[n]} + u^A_{a[n]} \\ &\quad + u^{BM}_{b[n],m[n]} + u^{BA}_{b[n],a[n]} + u^{MA}_{m[n],a[n]} \\ &\quad + u^{IM}_{b[n],i[n],m[n]} + u^{IA}_{b[n],i[n],a[n]} + u^{BMA}_{b[n],m[n],a[n]}, \end{aligned}$$

where every random effect is independently distributed as $u^F_\ell \sim \mathcal{N}(0, \sigma^2_F)$ for facet or interaction F. On the latent logistic scale, the observation-level residual variance is $\pi^2/3$. The reference fit contains $N = 29{,}923$ binary observations, 54 models, 13 scaffolds, nine benchmarks, and 1,117 benchmark-specific tasks.

The central object of measurement is the model. For a fixed evaluation design D, the relative generalizability coefficient has the schematic form from Equation 1

$$\rho^2_D = \frac{\sigma^2_M}{\sigma^2_M + \sigma^2_{\delta,D}}, \tag{16}$$

where $\sigma^2_{\delta,D}$ contains only those components that change model contrasts under repeated realizations of D, divided by their effective replication counts. Components that shift all models equally under a shared condition do not enter the relative error variance.

Cancellation of common condition effects. For two models m and m' evaluated under the same benchmark and scaffold,

$$\eta_{bima} - \eta_{bim'a}$$

eliminates the benchmark main effect u^B_b, scaffold main effect u^A_a, benchmark–scaffold effect u^{BA}_{ba}, task–scaffold effect $u^{IA[B]}_{bia}$, and task main effect u^I_{bi}. Thus, these components can affect absolute scores but not the ordering of models evaluated under identical conditions. This distinction is important below: instability in σ^2_A or σ^2_{BA} need not imply instability in relative model reliability.

All posterior chains for the reference model mixed adequately: the reported variance parameters had $\widehat{R} = 1.00$, with bulk effective sample sizes between 1,559 and 3,743. The following analyses address robustness to the data and specification rather than only Monte Carlo convergence.

E.2 Sensitivity of model rankings

Standard leaderboards rank systems by empirical mean accuracy,[12]

$$\bar{y}_{bm} = \frac{1}{N_{bm}} \sum_{n:\, b[n]=b, m[n]=m} y_n,$$

possibly aggregating over different scaffolds and nonidentical task sets. Such rankings treat all observed variation as evidence about model capability. The mixed models instead estimate latent model effects after separating task difficulty, scaffold effects, and their interactions.

[12] e.g., for HAL see "Accuracy" at `https://hal.cs.princeton.edu/reliability/` for HELM see "mean score" at `https://crfm.stanford.edu/helm/`, etc.

Table 9: Association between observed rankings and latent-effect rankings. "Full" uses the ecosystem model (Eq. 6); "Per-benchmark" uses an independently fitted model for each benchmark (Eq. 2)

Estimator	Metric	AssistantBench	CORE-Hard	GAIA	Mind2Web	SciCode	ScienceAgentBench	SWE-mini	τ-bench	USACO
Full	dCor_n^2	.404	.899	.780	.872	.703	.651	.833	.828	.430
Per-benchmark	dCor_n^2	-.005	.107	.046	.039	.013	.007	.046	.051	.032
Full	Kendall τ	.478	.859	.771	.744	.731	.582	.764	.787	.590
Per-benchmark	Kendall τ	.081	.248	.196	.151	.128	.101	.167	.180	.142
Full	Spearman ρ	.619	.961	.900	.896	.867	.795	.916	.930	.731
Per-benchmark	Spearman ρ	.111	.340	.271	.211	.171	.137	.236	.255	.199

For each benchmark, we formed rankings from posterior summaries of the relevant latent model effects. We considered two estimators:

1. **Full-model latent ranking**, obtained from the joint model in Equation 6, which partially pools information through models and scaffolds appearing elsewhere in the benchmark ecosystem.
2. **Per-benchmark latent ranking**, obtained by fitting, within each benchmark,

$$\mathrm{logit}(p_{ima}) = \mu + u_i^I + u_m^M + u_a^A + u_{im}^{IM} + u_{ia}^{IA} + u_{ma}^{MA}.$$

   This estimator uses only the information available within one benchmark.

These posterior conditional means are the Bayesian analogue of best linear unbiased predictors (BLUPs). We compared their induced rankings with the rankings from observed mean scores using Kendall's τ, Spearman's ρ, and bias-corrected squared distance correlation dCor_n^2 computed on rank vectors. Kendall's τ measures pairwise ordering agreement; Spearman's ρ measures monotone rank association; and corrected squared distance correlation can detect more general dependence between the rank assignments.

Two findings are salient. First, the jointly estimated rankings retain substantial agreement with observed leaderboards for CORE-Bench Hard, GAIA, Online-Mind2Web, SciCode, SWE-bench Verified Mini, and τ-Bench Airline. For example, full-model Spearman correlations are at least 0.85 on these benchmarks. Their empirical rankings therefore contain recoverable cross-system signal even after nuisance variation is separated.

Second, the independently fitted per-benchmark models exhibit weaker correspondence with raw rankings on every benchmark. This result is not evidence that the per-benchmark models are necessarily "less accurate." Instead, it exposes an information and estimand mismatch. Within one sparse benchmark, model effects must be separated from model–scaffold and task–model interactions using few connected observations. Strong posterior shrinkage is therefore appropriate, but it compresses distinctions that empirical average accuracy treats as model signal. Moreover, the raw ranking may combine deployable model–scaffold performance, whereas the latent model ranking intentionally removes scaffold-specific contributions. The joint model can recover more stable model effects because shared models and scaffolds connect otherwise isolated benchmark-specific designs. In other words, the combined model, after accounting for sources of variation, provides more stable and reliable estimates by taking advantage of shared facet variation.

Accordingly, rank correlation with the observed leaderboard is a sensitivity diagnostic, not a ground-truth accuracy measure. High agreement indicates that adjustment for known nuisance facets preserves the reported ordering. Low agreement indicates that the ordering depends on whether task and scaffold variation are treated as capability or as measurement error. It does not establish which ordering is externally valid without an independent criterion.

E.3 Pointwise predictive stability with PSIS-LOO

We first assessed whether individual observations exert disproportionate influence on the posterior predictive distribution. Exact leave-one-out cross-validation would require N refits. We instead used Pareto-smoothed importance-sampling leave-one-out cross-validation (PSIS-LOO), with moment matching for observations having unstable importance ratios (Vehtari et al., 2017; Paananen et al.,

Table 10: PSIS-LOO diagnostics for the full Bayesian logistic mixed model.

Quantity	Estimate	Standard error
$\mathrm{elpd}_{\mathrm{loo}}$	$-11{,}217.1$	94.2
p_{loo}	3,032.0	32.0
LOOIC	22,434.2	188.5
Pareto diagnostic	**Count**	**Percentage**
$\widehat{k} \leq 0.7$	29,902	$> 99.9\%$
$0.7 < \widehat{k} \leq 1$	21	$< 0.1\%$
$\widehat{k} > 1$	0	0.0%

2021; Vehtari et al., 2024). For observation n, PSIS approximates

$$p(y_n \mid y_{-n}) = \int p(y_n \mid \theta)\, p(\theta \mid y_{-n})\, d\theta$$

using draws from the full posterior $p(\theta \mid y)$. The generalized Pareto shape diagnostic $\widehat{k}_n$ measures the tail behavior of the resulting importance ratios. Values $\widehat{k}_n \leq 0.7$ generally indicate a reliable approximation, whereas larger values identify observations for which deleting the point substantially changes its posterior predictive distribution.

The diagnostics in Table 10 are favorable for a model of this complexity. The effective number of parameters is substantially smaller than both N and the nominal number of coefficients induced by the random effects, demonstrating strong regularization through partial pooling. No observation has $\widehat{k} > 1$, and only 21 of 29,923 observations have $\widehat{k} > 0.7$.

The few warnings are explained by the connectivity of the design rather than by broad model failure. Approximately 32% of observed benchmark–task–model groups are singletons, a known point of sensitivity for PSIS-LOO (Bindoff, 2026). Of the 21 observations with $\widehat{k} > 0.7$, 19 (90.5%) belong to such singleton groups. The remaining two form the only observations in their group: SciCode task 14 evaluated with GPT-4.1 under two scaffolds.

This behavior follows directly from the pointwise LOO target. If y_n is the only observation informing a random-effect level u_ℓ, deletion of y_n makes the leave-one-out distribution of u_ℓ approximately prior-predictive. The full-data posterior, by contrast, has adapted u_ℓ to y_n. Importance sampling must therefore bridge two meaningfully different distributions, producing a heavy-tailed importance ratio. A large $\widehat{k}_n$ in this setting primarily says that the outcome of an isolated, data-defined group cannot be predicted after removing its only observation. It does not, by itself, imply that population-level variance components are determined by that point.

E.3.1 Observation-level random-effect ablation.

As an additional check, we augmented Equation 6 with

$$u^{IMA}_{b,i,m,a} \sim \mathcal{N}(0, \sigma^2_{IMA}).$$

Because benchmark–task–model–scaffold cells generally lack within-cell replication in the HAL dataset, this term acts as an observation-level random effect (OLRE) (Harrison, 2015). In a Bernoulli-logit model, it competes with the fixed latent logistic residual rather than identifying a conventionally replicated three-way interaction. Including it did not materially change the reliability estimates: it absorbed variation previously treated as observation-level noise, and its contribution is attenuated by task replication in the D-study. The principal generalizability conclusions are therefore not an artifact of omitting the highest-order cell term.

Importantly, this additional term should be used for studies where output stability is of interest where benchmark–task–model–scaffold levels have multiple runs. This variation can support measuring the stochasticity of task scores under the same conditions.

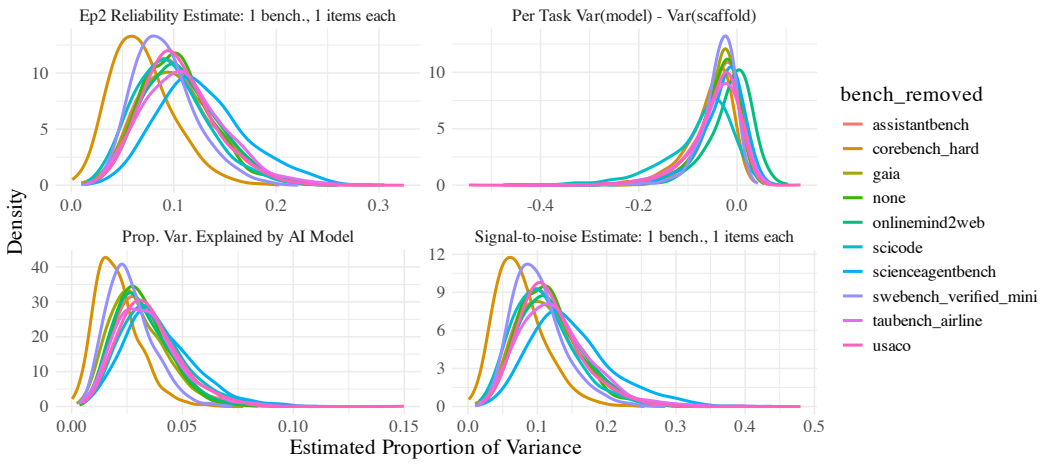


Figure 6: Posterior distributions for benchmark-level LOO ablations for the paper's quantities of interest

E.4 Leave-one-benchmark-out sensitivity

Pointwise LOO evaluates interpolation within the observed ecosystem. A stronger test removes an entire benchmark and therefore deletes all of its tasks, score distribution, and benchmark-specific interactions. For each $b^\star \in \mathcal{B}$, we refitted the complete variance decomposition to

$$\mathcal{D}_{-b^\star} = \{y_n : b[n] \neq b^\star\}$$

and recomputed the posterior variance decomposition and D-study reliability curves. This procedure asks whether conclusions about the benchmark battery are broadly distributed across benchmarks or are driven by a single test.

The reliability estimations, proportions of variance attributable to models, and the principal interaction facets were generally stable across these refits as shown in Fig. 6. Sensitivity nevertheless depended on which benchmark was withheld. Removing CORE-Bench Hard caused the largest reduction in estimated reliability. This is substantively expected: CORE-Bench Hard supplies the most runs, spans the widest range of observed accuracies, and has the highest estimated signal-to-noise ratio among the included benchmarks. It therefore contributes unusually strong information both for distinguishing models and for connecting model performance to the rest of the ecosystem.

Without CORE-Bench Hard, the posterior signal-to-noise ratio from the D-study no longer reaches the prespecified limit-of-detection band. This finding should not be read merely as "more observations are better." A large but noisy benchmark can contribute little to the reliability of a battery. CORE-Bench Hard is influential because it combines replication with discrimination. Thus, increasing the number of benchmarks is not equivalent to increasing effective measurement information.

More generally, for a battery aggregating conditionally independent benchmark-level measurements with signal S_b and error E_b, the information supplied by benchmark b is governed by its discrimination relative to error, not simply its task count. Although the fully crossed design includes interactions and is more complicated than this schematic case, the same principle applies: benchmarks with substantial model variation and controlled task- and harness-dependent error contribute disproportionately to stable ecosystem-level rankings.

This ablation also clarifies the intended scope of generalization. The leave-one-benchmark-out results do not claim that the fitted model can predict an arbitrary future benchmark with no shared structure. Rather, they test whether the estimated variance decomposition and design recommendations survive removal of one observed measurement instrument. The sensitivity to CORE-Bench Hard indicates that the present nine-benchmark ecosystem contains limited redundancy at the high-signal end.

Table 11: Leave-one-benchmark–scaffold-out proportions of variance. SE is the standard error across refits, and CV is the coefficient of variation.

Facet	Mean	Median	Min.	Max.	SD	SE	CV
Residual	.237	.235	.226	.267	.008	.002	.035
Scaffold (A)	.020	.019	.008	.037	.007	.002	.338
Benchmark (B)	.258	.259	.229	.298	.015	.003	.060
Benchmark–scaffold (BA)	.026	.028	.000	.035	.009	.002	.345
Benchmark–model (BM)	.020	.020	.009	.031	.004	.001	.208
Benchmark–model–scaffold (BMA)	.014	.013	.010	.019	.003	.001	.180
Task within benchmark (I)	.328	.329	.282	.365	.015	.003	.047
Task–scaffold (IA)	.052	.053	.028	.066	.009	.002	.168
Task–model (IM)	.003	.004	.000	.006	.001	$< .001$	.426
Model (M)	.033	.033	.026	.044	.004	.001	.110
Model–scaffold (MA)	.008	.009	.003	.012	.002	$< .001$	.249

E.5 Leave-one-benchmark–scaffold-out sensitivity

Because scaffold coverage is also unbalanced, we conducted a complementary sensitivity analysis that removed one observed benchmark–scaffold combination at a time and re-estimated the full decomposition. Table 11 summarizes the resulting frequentist proportions of total variance. This ablation is especially stringent for benchmark-specific scaffolds, whose removal can eliminate nearly all direct information about a scaffold level.

Table 12: Leave-one-LLM-out proportions of variance. SD is the standard deviation across refits, SE is the standard error across refits, and CV is the coefficient of variation.

Facet	Mean	Median	Min.	Max.	SD	SE	CV
Residual	0.236	0.236	0.231	0.241	0.002	0	0.008
Scaffold (A)	0.020	0.020	0.016	0.029	0.002	0	0.098
Benchmark (B)	0.259	0.259	0.252	0.268	0.002	0	0.009
Benchmark–scaffold (BA)	0.026	0.026	0.022	0.030	0.001	0	0.056
Benchmark–model (BM)	0.020	0.020	0.015	0.022	0.001	0	0.065
Benchmark–model–scaffold (BMA)	0.014	0.014	0.008	0.016	0.001	0	0.092
Task within benchmark (I)	0.328	0.328	0.320	0.334	0.002	0	0.006
Task–scaffold (IA)	0.052	0.052	0.049	0.053	0.001	0	0.018
Task–model (IM)	0.004	0.004	0.001	0.004	0.001	0	0.166
Model (M)	0.033	0.033	0.028	0.036	0.002	0	0.048
Model–scaffold (MA)	0.009	0.009	0.006	0.012	0.001	0	0.107

The dominant components are stable. Task difficulty accounts for approximately 32.8% of variance across refits, benchmark differences for 25.8%, and residual variation for 23.7%. Their coefficients of variation are only 0.047, 0.060, and 0.035, respectively. The model component is also stable in absolute terms, ranging from 0.026 to 0.044.

The largest relative variation occurs for the scaffold main effect, the benchmark–scaffold interaction, and the task–model interaction. These cases require different interpretations. The first two are weakly identified because most scaffolds occur in only one benchmark; removing a benchmark–scaffold cell can therefore remove much of the relevant connectivity. However, common scaffold and benchmark–scaffold shifts cancel from same-condition model contrasts and consequently do not enter the relative model reliability estimates used in the primary D-study.

The task–model component has the largest coefficient of variation (0.426), but its estimated proportion is always between 0 and 0.006. Its high relative variability is therefore primarily a small-denominator effect: even its maximum is smaller than the typical contribution of every other reported component. Coefficients of variation should not be interpreted without the corresponding absolute scale. Additionally, both this component and the second largest CV value, the benchmark–scaffold component, are the only components that contain minimum values equal to zero; these sensitivities, paired with

extreme ablation minima, are likely elevated due to the nature of frequentist estimations, which can result in singular values in highly unbalanced designs (see Appendix H.3). These estimates may be more stable under a full Bayesian ablation where CV estimates would be less likely to decrease the mean in the denominator.

The model-related components that directly govern ranking stability remain comparatively well behaved. The model proportion has CV 0.110, while the benchmark–model and model–scaffold components remain small across all deletions. Hence, no individual benchmark–scaffold cell appears to create the principal conclusion that bare-model rankings are less reliable than rankings would appear under a decomposition that treats all observed model–scaffold performance as signal. However, it is worth noting that removing AssistantBench during the LOO ablation disconnects the OnlineMind2Web scaffold from the broader network, which means that, for that connection, identifiability conditions found in Proposition F.3 would not be complete. Nevertheless, the estimates remain stable, even with this disconnection.

Importantly, the contribution to overall reliability of CORE-Bench Hard discussed in § E.4 can be narrowed further to the contribution of the CORE Agent, which shows the most between-LLM discrimination on its benchmark. Removing this particular benchmark–scaffold accounts for the large increase in uncertainty across the pooled model. As this scaffold is less connected than the HAL Generalist scaffold, this reaffirms that having a strong discriminating signal is critical to pooled reliability, potentially more than large connectivity.

E.6 Leave-one-LLM-out sensitivity

We next test whether the estimated measurement structure is driven by a single LLM and whether the variance decomposition generalizes to new LLMs. For each LLM model $m^\star \in \mathcal{M}$, we removed all of its observations,

$$\mathcal{D}_{-m^\star} = \{y_n : m[n] \neq m^\star\},$$

and refit the variance-decomposition model. This is a stronger perturbation than deleting one response: it simultaneously removes a model main-effect level and every observed benchmark–model, model–scaffold, task–model, and benchmark–model–scaffold cell involving that LLM. Because evaluation coverage differs substantially across models—only nine of the 54 models appear in every benchmark—the ablation also tests sensitivity to the most highly connected models in the design. The LOO ablations for this and those of Appendix E.5 were conducted in a frequentist framework to reduce practical computation costs of hundreds of Bayesian refittings (see Appendix C.6).

Table 12 shows that the decomposition is remarkably insensitive to the removal of any LLM. The three largest components are almost invariant: task-within-benchmark variance has mean proportion 0.328, range 0.320–0.334, and CV 0.006; residual variance has mean 0.236, range 0.231–0.241, and CV 0.008; and benchmark variance has mean 0.259, range 0.252–0.268, and CV 0.009. Thus, the conclusion that variation is dominated by differences among tasks and benchmarks, together with substantial unexplained response variation, is not attributable to the performance profile of a particular model.

More importantly for relative reliability, the model variance is also stable. Its proportion remains between 0.028 and 0.036, with mean 0.033 and CV 0.048. The benchmark–model component ranges from 0.015 to 0.022, while the benchmark–model–scaffold component ranges from 0.008 to 0.016. Removing even a highly connected or unusually capable LLM therefore does not qualitatively change the estimated amount of systematic between-model variation or the extent to which model performance depends on the benchmark and scaffold.

As in the benchmark–scaffold ablation, the largest relative variability occurs in small components. The task–model interaction has CV 0.166, but its variance share is only 0.001–0.004. The model–scaffold interaction has CV 0.107 and remains between 0.006 and 0.012. Their relative sensitivity should consequently not be confused with a large contribution to total variance. In absolute terms, deletion-induced changes in both components are small.

The leave-one-model results are also more stable than the leave-one-benchmark results. This asymmetry reflects the structure of the available evidence. Each LLM contributes another sample from the population of systems, and its outcomes are partially pooled with those of the remaining 53 models. By contrast, removing a benchmark deletes an entire measurement instrument, all of its unique tasks,

and its characteristic signal-to-noise ratio. The ecosystem therefore has greater redundancy across models than across high-quality benchmarks. Adding another LLM primarily improves estimation of the distribution of model capability, whereas adding a discriminating benchmark can alter the quality of the measurement battery itself.

Implication for reliability. For a fixed D-study design, relative reliability depends on the model variance and on model-dependent error components such as BM, MA, IM, and BMA, after scaling by their effective replication counts. The stability of these components under model deletion implies that the reported reliability conclusions are not generated by one extreme or unusually well-connected LLM. Nevertheless, this ablation evaluates *influence on the population variance decomposition*, not the ability to predict the performance or rank of a previously unseen model. Generalization to a new LLM additionally requires that it be exchangeable with the sampled model population and evaluated on conditions that connect it to the existing design.

No single LLM acts as a leverage point for the principal variance decomposition. The greater sensitivity to removing an informative benchmark than to removing an LLM reinforces a central design recommendation: once a reasonably diverse model sample has been obtained, additional evaluation resources may yield greater reliability gains by improving benchmark quality, task replication, and cross-scaffold connectivity than by adding sparsely evaluated models.

E.7 Methodological Robustness

We also estimate each model using five different approaches. A separate Appendix H explains and discusses each method and reports the full results.

E.8 Rank stability across posterior

Ranking models within each draw yields posterior rank distributions and pairwise ordering probabilities $\Pr(\theta_{mb} > \theta_{m'b} \mid y)$. These quantities distinguish an estimated ordering from evidence that two models are meaningfully distinguishable. We compare posterior and published rankings using Spearman correlation and ranked unbiased squared distance correlation, dCor_n^2; the latter remains informative in the presence of extensive ties.

We compare posterior ranks with published ranks using Spearman correlation and ranked unbiased squared distance correlation, dCor_n^2 (Table 13). The latter is useful for benchmarks with extensive ties and provides a direct diagnostic of whether estimated capability ranking is statistically associated with the reported ordering.

Posterior capability estimates differ from ranks obtained by sorting raw percent-correct scores. Raw scores credit the model for every condition with which it happens to be paired. The pooled decomposition instead estimates the portion of performance that persists after averaging over the specified benchmark and scaffold universes.

Table 13: Median correlation of reported and estimated ranks across posterior draws

Corr.	AssistantBench	CORE-Bench	GAIA	Mind2web	SciCode	Sci.AgentBench	SWEbench	τ-bench	USACO
Spearman	0.442	0.824	0.79	0.621	0.662	0.611	0.808	0.683	0.692
dCor_n^2	0.141	0.64	0.576	0.33	0.36	0.323	0.633	0.408	0.395

E.9 Implications for benchmark design

These checks support five practical conclusions.

1. **Sparse cells principally limit local prediction.** PSIS warnings are almost entirely confined to singleton or doubleton interaction levels. The model cannot predict an isolated cell after its sole observation is removed, but the population-level variance decomposition is stable to these observations.
2. **Partial pooling is necessary for ecosystem-level ranking.** Per-benchmark data are generally insufficient to cleanly distinguish model capability from scaffold and task interactions.

Cross-benchmark connectivity substantially stabilizes latent model rankings, although the resulting rankings remain uncertain on low-signal benchmarks.

3. **Benchmark quality is not interchangeable with benchmark quantity.** The leave-one-benchmark-out analysis identifies CORE-Bench Hard as a high-information anchor. A useful test battery should include multiple independently constructed benchmarks with high discrimination and controlled error, rather than merely adding more noisy tasks or near-duplicate benchmarks.
4. **Absolute-score instability and relative-rank instability are distinct.** Scaffold and benchmark–scaffold effects can materially shift reported accuracies while canceling from same-condition model comparisons. Benchmark reports should therefore state whether reliability concerns absolute deployment performance, model ranking, or model–scaffold system ranking; these are different objects of measurement and induce different error terms.
5. **Representative LLM panel improves benchmark diagnostics.** once a reasonably diverse model sample has been obtained, additional evaluation resources may yield greater reliability gains by improving benchmark quality, task replication, and cross-scaffold connectivity than by adding sparsely evaluated models.

The robustness analyses do not imply that the sparse design is harmless. Rather, they localize its consequences. The main variance and reliability conclusions are not driven by a handful of observations or benchmark–scaffold cells, but the precision of individual rankings remains strongly dependent on cross-benchmark connectivity and on the inclusion of at least one high-signal measurement instrument. This distinction is essential for designing future agentic benchmark batteries: additional evaluation should be allocated to conditions that improve connectivity and reduce model-dependent error, not only to increasing the nominal number of tasks or sparsely connecting models.

F Identifiability and connectivity of the leaderboard decomposition

The leaderboard data are sparse by construction: models are submitted selectively, scaffolds are not used uniformly, and tasks are unique to benchmarks. Such missingness is common in public leaderboards (Singh et al., 2026). A fully crossed experiment would simplify estimation, but it is not necessary for the questions studied here. What is required is sufficient overlap to distinguish persistent model variation from variation associated with benchmarks, scaffolds, and their interactions.

This appendix establishes that the observed design provides the connectivity needed for those contrasts. We distinguish three concepts that are often conflated:

1. **Connectivity:** whether observed model, benchmark, and scaffold levels belong to a common comparison network.
2. **Structural identifiability:** whether distinct variance components imply distinct distributions over the observed responses.
3. **Practical estimability:** whether the finite data determine those components precisely.

Connectivity and structural identifiability are design properties; practical estimability also depends on sample size, outcome variation, and prior regularization. Our Bayesian estimator can produce a proper posterior for a weakly informed component, but a proper posterior alone is not evidence that the component is strongly data-identified. We therefore use posterior uncertainty and sensitivity analyses, rather than existence of an estimate, to characterize practical estimability.

F.1 Incidence-Graph Connectivity and Identifiability

Represent the observed design as a multipartite incidence graph whose vertices are benchmarks, models, and scaffolds, with edges induced by observed evaluations. Shared models and scaffolds connect benchmark-specific observations and support estimation of common variance components. Contrasts within a connected component are informed by observed paths; contrasts across disconnected components are not identified without additional assumptions. We report component membership, articulation vertices, benchmark degrees, and the change in connectivity produced by removing each benchmark. Posterior regularization stabilizes weakly supported components but does not create evidence for disconnected contrasts.

F.2 Observed incidence structure

The full dataset contains nine benchmarks, 54 LLMs, and 13 scaffolds. Nine models appear on every benchmark and span at least 69% of the scaffold set. One scaffold is shared by eight benchmarks. AssistantBench connects the remaining benchmark through an additional shared scaffold, and every other benchmark contains at least one additional benchmark-specific scaffold. Models and scaffolds are consequently neither fully crossed nor evenly replicated.

Let

$$\Omega = \{(b, i, m, a) : y_{bima} \text{ is observed}\}$$

denote the observed response cells. The pooled decomposition is equation 6:

$$\begin{aligned}\eta_{bima} = \beta_0 + u_b^{(B)} + u_{bi}^{(I[B])} + u_m^{(M)} + u_a^{(A)} + u_{bm}^{(BM)} + u_{ba}^{(BA)} + u_{ma}^{(MA)} \\ + u_{bim}^{(IM[B])} + u_{bia}^{(IA[B])} + u_{bma}^{(BMA)}, \qquad (b, i, m, a) \in \Omega.\end{aligned}$$

Each random effect has mean zero and a component-specific variance. The Bernoulli–logit likelihood fixes the latent scale through the standard logistic residual variance $\pi^2/3$.

It is useful to represent Ω as a multipartite incidence graph. Let

$$G = (V, E), \qquad V = \mathcal{B} \cup \mathcal{M} \cup \mathcal{A},$$

where an edge joins two levels when they co-occur in at least one observed response. For example, m–b is an edge if model m is evaluated on benchmark b, and a–b is an edge if scaffold a is used on benchmark b. Items need not connect across benchmarks because they are intentionally nested within benchmark.

The nine models observed on every benchmark form model-side anchors. The scaffold used on eight benchmarks forms a scaffold-side anchor, while the second shared scaffold connects AssistantBench to the rest of the graph. Benchmark-specific scaffolds are leaves or local branches attached to this connected core. Thus, all benchmarks and their associated observations belong to one comparison network.

Lemma 1 (connected additive contrasts).

Lemma F.1 (connected additive contrasts)**.** *Consider an additive model on an observed incidence graph,*

$$\eta_{ma} = \mu + \alpha_m + \gamma_a,$$

with one centering constraint per facet. If the model–scaffold incidence graph is connected, all estimable contrasts $\alpha_m - \alpha_{m'}$ and $\gamma_a - \gamma_{a'}$ are identified.

Proof. Suppose two parameterizations produce the same linear predictor on every observed edge. Their differences satisfy $\Delta\alpha_m + \Delta\gamma_a = 0$ on each edge. Along any path in a connected bipartite graph, these equalities imply that all model differences equal a common constant and all scaffold differences equal its negative. Centering removes this remaining additive degree of freedom. □

Lemma F.1 concerns fixed additive effects, whereas Equation 6 uses random effects and interactions. It nevertheless provides the relevant intuition: an observation from a model or scaffold that is disconnected from the remainder of the leaderboard cannot support a common ranking. Shared models and scaffolds create paths along which relative effects can be compared.

F.3 Identifiability of random-effect variances

For observations ordered as a vector, write the latent linear predictor as

$$\boldsymbol{\eta} = \beta_0 \mathbf{1} + \sum_{k=1}^{K} Z_k \mathbf{u}_k, \qquad \mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 I),$$

where Z_k is the incidence matrix for component k. On the latent scale, the random effects induce covariance

$$\mathrm{Cov}(\boldsymbol{\eta}) = \sum_{k=1}^{K} \sigma_k^2 K_k, \qquad K_k = Z_k Z_k^\top. \tag{17}$$

Table 14: Observed comparisons supporting the pooled variance decomposition. "Separation" identifies the principal pair of components distinguished by each overlap pattern.

Separation	Required comparison	Support in the observed design
M versus BM	The same model appears on multiple benchmarks.	Nine models appear on all nine benchmarks; additional models provide partial cross-benchmark replication.
A versus BA	The same scaffold appears on multiple benchmarks.	One scaffold spans eight benchmarks; an additional shared scaffold connects AssistantBench.
MA versus BMA	The same model–scaffold pair recurs across benchmarks.	Cross-benchmark models evaluated through shared scaffolds create repeated model–scaffold pairs.
M versus MA	Models are observed under multiple scaffolds, with overlap across models.	The cross-benchmark models span at least nine of the 13 scaffolds, and shared scaffolds provide common comparison conditions.
BM versus BMA	Within a benchmark, models are observed under more than one scaffold.	Each benchmark contains shared or locally replicated scaffold conditions in addition to benchmark-specific scaffolds.
$I[B]$ versus $IM[B]$	Each item is attempted by multiple models.	Benchmark items are repeatedly scored across leaderboard models.
$I[B]$ versus $IA[B]$	Each item is attempted under multiple scaffold conditions.	Scaffold replication within benchmarks provides item–scaffold contrasts where observed.

The entries of K_k indicate which pairs of observations share a level of facet or interaction k. For example, two observations share the model kernel K_M when they use the same model, and share K_{BM} only when they use both the same benchmark and model.

Proposition F.2 (variance-component criterion)**.** *A collection of latent variance components* $\{\sigma_k^2 : k \in \mathcal{K}\}$ *is structurally identifiable from the observed design if its covariance kernels* $\{K_k : k \in \mathcal{K}\}$*, restricted to* Ω*, are linearly independent after removing components that are deterministically confounded with the terminal residual.*

Proof. If the restricted kernels are linearly independent, equality of two induced covariance matrices implies

$$\sum_{k \in \mathcal{K}} (\sigma_k^2 - \widetilde{\sigma}_k^2) K_k = 0,$$

which has only the trivial solution $\sigma_k^2 = \widetilde{\sigma}_k^2$ for every k. If the kernels are linearly dependent, a nonzero perturbation of their coefficients leaves the induced covariance unchanged, so the corresponding components cannot be separated from the observed design. □

For a Bernoulli GLMM, Proposition F.2 is most directly interpreted as a design criterion on the latent scale. The nonlinear likelihood can affect the amount of information, especially under floor effects, but it cannot create distinctions absent from the incidence matrices.

A practical interpretation is the following: to distinguish two variance components, the design must contain pairs of observations that share the grouping represented by one component without always sharing the grouping represented by the other. Complete crossing is sufficient for this condition but is not necessary.

F.4 How the observed design separates the required components

Table 14 summarizes the principal replication patterns. These conditions concern variance components, not estimation of every individual random-effect realization.

F.4.1 MODEL AND BENCHMARK–MODEL VARIATION

The distinction between persistent model capability and benchmark-specific performance is central to the leaderboard reliability coefficient. If every model appeared on only one benchmark, then $u_m^{(M)}$ and $u_{bm}^{(BM)}$ would be inseparable: "model" would be nested within benchmark. That failure does not occur here. Nine models appear on every benchmark, producing observations that share M while differing in BM.

Consequently, the model kernel K_M and benchmark–model kernel K_{BM} have different support. The former links observations from the same model across benchmarks; the latter links them only within a benchmark. This overlap identifies the contrast between globally persistent model variation σ_M^2 and benchmark-conditioned model variation σ_{BM}^2.

F.4.2 SCAFFOLD AND BENCHMARK–SCAFFOLD VARIATION

Benchmark-specific scaffolds alone would not distinguish a scaffold main effect from a benchmark–scaffold interaction. For a scaffold used on exactly one benchmark, its A and BA columns coincide. The design avoids complete confounding because one scaffold spans eight benchmarks and a second shared scaffold connects AssistantBench. Observations using a shared scaffold have the same A level but different BA levels, providing the comparisons required to distinguish σ_A^2 from σ_{BA}^2.

Benchmark-specific scaffolds remain less individually informed than shared scaffolds. Their effects are estimated through the exchangeability assumptions of the hierarchical model, with uncertainty propagated into the posterior. The shared scaffolds identify the population-level separation; partial pooling regularizes levels with little direct replication.

F.4.3 MODEL–SCAFFOLD AND BENCHMARK–MODEL–SCAFFOLD VARIATION

Separating MA from BMA requires recurrence of a model–scaffold pair across benchmarks. The cross-benchmark models and shared scaffolds provide such recurrence: observations can share a model and scaffold while differing in benchmark. Within-benchmark scaffold variation then supplies the complementary comparisons needed to estimate how model–scaffold compatibility changes by benchmark.

This component is more weakly informed than the model and benchmark–model components because repeated model–scaffold pairs are less frequent than repeated models. Our estimands therefore integrate over its posterior uncertainty. The leave-one-benchmark-out and alternative-estimator analyses in Appendix E assess whether substantive conclusions depend on a small number of these bridges.

F.4.4 ITEM-INDEXED VARIATION

Items are nested within benchmarks and are not expected to recur across benchmarks. Their main effects are identified because the same item is attempted by multiple models and scaffold conditions. Item–model variation is informed when a benchmark item is attempted by multiple models, while item–scaffold variation is informed by scaffold replication within benchmark.

No assumption is made that an item from one benchmark is exchangeable with an identically labeled item from another benchmark. Pooling occurs at the variance-component level: the model estimates the typical magnitude of item-conditioned interactions across the observed benchmark panel.

F.5 WHAT IS NOT SEPARATELY IDENTIFIABLE

The design does not contain independent repeated executions of every (b, i, m, a) cell. A four-way benchmark–item–model–scaffold interaction is therefore observationally confounded with cell-level execution variation and the Bernoulli residual for most of such interactions. We combine these sources into the terminal component

$$\sigma_{BIMA,e}^2.$$

This is not a limitation for the reported relative reliability coefficients: all unresolved cell-specific variation belongs in the error term because it can change model ordering across repeated evaluation

conditions. Separating execution stochasticity, grading instability, and four-way interaction would require repeated runs under the same recorded cell and, ideally, explicit grading replicates.

More generally, the present design does not support:

- an unregularized estimate for every unobserved model–scaffold cell;
- precise scaffold-specific effects for scaffolds used on only one benchmark;
- causal claims about replacing one scaffold with another; or
- generalization to scaffold or benchmark populations unrelated to the connected evaluation panel.

These quantities are unnecessary for our research questions. We require population-level variance components and scaffold-marginalized model contrasts, not predictions for every missing factorial cell.

F.6 Identifiability of the reliability estimands

The principal leaderboard coefficient is Equation 7:

$$E\rho_M^2(n_b, n_i, n_a) = \frac{\sigma_M^2}{\sigma_M^2 + \frac{\sigma_{BM}^2}{n_b} + \frac{\sigma_{MA}^2}{n_a} + \frac{\sigma_{BMA}^2}{n_b n_a} + \frac{\sigma_{IM[B]}^2}{n_b n_i} + \frac{\sigma_{BIMA,e}^2}{n_b n_i n_a}}.$$

This coefficient depends on a small set of variance components and their sums, not on every random effect in Equation 6. The design requirements are correspondingly weaker than those needed to reconstruct the complete model–benchmark–scaffold response tensor.

Proposition F.3 (sufficiency for the leaderboard estimand)**.** *Suppose that:*

(i) the benchmark–model incidence graph is connected and contains models observed on multiple benchmarks;
(ii) the benchmark–scaffold incidence graph is connected and contains scaffolds observed on multiple benchmarks;
(iii) at least some model–scaffold pairs recur across benchmarks; and
(iv) benchmark items are attempted by multiple models and scaffolds.

Then the covariance kernels corresponding to M, BM, MA, BMA, and $IM[B]$ have distinct observed replication patterns. The variance combinations entering Equation 7 are therefore structurally distinguishable, up to the explicitly combined terminal component $\sigma_{BIMA,e}^2$.

Justification Conditions (i)–(iv) provide observation pairs that share, respectively: M but not BM; A but not BA; MA but not BMA; and $IM[B]$ and not only $I[B]$. These yield distinct covariance kernels under Proposition F.2. The cell-specific remainder is intentionally aggregated because no additional replication distinguishes its constituents. □

The observed leaderboard satisfies these conditions. In particular, the nine cross-benchmark models identify persistent versus benchmark-conditioned model variation, while shared scaffolds and repeated model–scaffold pairs identify the scaffold-related components that determine the task-only reliability ceiling.

Corollary F.4 (identifiability of D-study ceilings)**.** *Under Proposition F.3, the task-only limit*

$$\lim_{n_i \to \infty} E\rho_M^2(n_b, n_i, n_a) = \frac{\sigma_M^2}{\sigma_M^2 + \sigma_{BM}^2/n_b + \sigma_{MA}^2/n_a + \sigma_{BMA}^2/(n_b n_a)}$$

is identified without separately decomposing terminal item-level residual variation.

Corollary F.4 is important for the paper's central design conclusion. The claim that adding tasks cannot eliminate benchmark- and scaffold-conditioned error does not depend on a precise decomposition of every item-level noise source. It depends on the persistent components whose identification is supported by cross-benchmark models and scaffolds.

F.7 Bayesian regularization and evidential scope

All variance components are estimated with proper, weakly informative priors. The posterior is therefore proper even when a component is weakly informed near the boundary. We use priors to

stabilize finite-sample estimation, not to substitute for disconnected comparisons. Three features constrain interpretation:

1. **Data-supported contrasts.** Model rankings are inferred through the connected observation graph. We do not report comparisons between disconnected components because none exist in the observed design.
2. **Posterior rather than plug-in reliability.** Equation 7 is evaluated within each posterior draw. Weak separation among related components therefore appears as uncertainty in reliability and its D-study projection.
3. **Sensitivity to bridges.** Leave-one-benchmark-out refits test whether conclusions rely on a particular benchmark or scaffold connection. Prior and estimator comparisons test whether weak components are driving the reported variance ordering. These analyses are reported in Appendix E.

The distinction between identifiability and precision is especially important for agent leaderboards. Sparse designs can be connected enough to estimate whether scaffold variation is non-negligible while still being too sparse to estimate its exact magnitude narrowly. Large credible intervals in this setting are not model failure; they reveal that the leaderboard contains too few common evaluation conditions to support precise claims.

F.8 Connectivity Summary

The full design is incomplete but connected. Nine models evaluated across every benchmark anchor the model scale; shared scaffolds connect the benchmark panel; and recurring model–scaffold pairs distinguish general scaffold compatibility from benchmark-specific compatibility. These overlaps are sufficient for the variance combinations underlying model-ranking reliability, signal-to-noise ratios, scaffold-versus-model comparisons, and D-study ceilings.

The design does not identify every possible cell effect, nor is that required. Unreplicated terminal variation is conservatively assigned to relative error, and uncertainty from sparsely observed scaffold interactions is propagated through the Bayesian posterior. The resulting claims are therefore appropriately scoped: they characterize the reliability supported by this connected leaderboard ecosystem, rather than asserting a complete factorial decomposition of all possible models, scaffolds, tasks, and benchmarks.

G Additional Tables and Figures

The coefficients of Table 1 isolate *where* a benchmark's signal lives and *which* facet dissipates it. Because each coefficient shares a common numerator–denominator structure (signal variance over signal plus generalizing error), differences across coefficients within a single benchmark are directly interpretable: they reveal how much of the apparent capability signal survives when we generalize over tasks, over scaffolds, or over both. Table 16 reports posterior means from the per-benchmark fit of Eq. 2.

Table 15: **Reliability depends on the object being ranked.** Reliability estimates for each benchmark under model and model–scaffold systems as measurement systems, including the task limit for models as the object of measurement.

Estimand	Assistant	CORE-Hard	GAIA	Mind2Web	SciCode	ScienceAgent	SWE-mini	τ-Airline	USACO
$\lim_{n_i\to\infty} E\rho^2_M$ (8)	0.210	0.892	0.325	0.153	0.840	0.683	0.552	0.669	0.377
$E\rho^2_M$ (3)	0.182	0.841	0.318	0.148	0.811	0.628	0.543	0.572	0.375
$E\rho^2_{MA}$ (4)	0.935	0.972	0.990	0.974	0.975	0.987	0.993	0.944	0.994

G.1 Repeated Task Subsampling

We confirm the cost reduction findings by repeatedly sampling k tasks per benchmark without replacement, using the same sampled tasks for all models to ensure a paired comparison. Within each subsample, scores are first averaged at the benchmark–model level and then averaged across benchmarks, giving each benchmark equal weight. The resulting model ranking is compared with

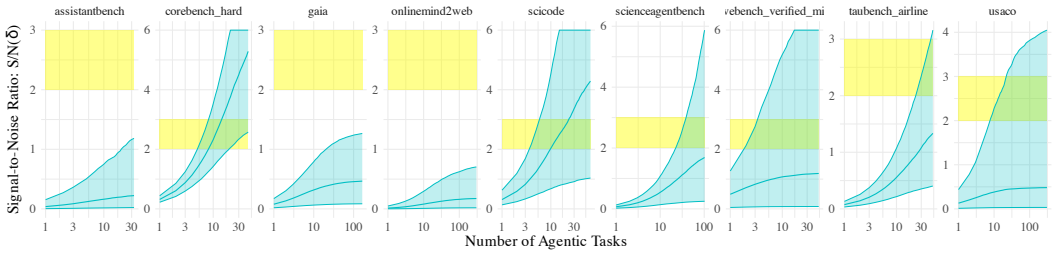


Figure 7: **Benchmarks differ sharply in ranking signal.** Posterior model-ranking signal-to-noise ratio as tasks are added. Shaded bands give external detection-limit reference values. Points and intervals are posterior medians and 68% standard error HDIs.

Table 16: Posterior-mean reliability and generalizability coefficients per benchmark. $E\rho^2_{M(i)}$ and $E\rho^2_{A(i)}$ are task reliabilities ($\sigma^2_o/(\sigma^2_o + \sigma^2_{io})$) for model and scaffold rankings; $\rho^{(b)}_{AA'}$ is inter-scaffold reliability (Eq. 5); $E\rho^2_{M(b)}$, $E\rho^2_{A(b)}$, and $E\rho^2_{MA(b)}$ are G-coefficients by object of measurement (Eqs. 3–4); the last column is the task-saturated model ceiling $\lim_{n_i\to\infty} E\rho^2_{M(b)}$ (Eq. 8).

Benchmark	$E\rho^2_{M(i)}$	$E\rho^2_{A(i)}$	$\rho^{(b)}_{AA'}$	$E\rho^2_{M(b)}$	$E\rho^2_{A(b)}$	$E\rho^2_{MA(b)}$	$E\rho^2_{M(b)}(\infty)$
scicode	0.602	0.534	0.746	0.736	0.516	0.971	0.765
gaia	0.263	0.705	0.333	0.328	0.579	**0.990**	0.334
taubench_airline	0.549	0.539	0.545	0.533	0.759	0.938	0.619
swebench_verified_mini	0.599	0.857	0.504	0.499	0.639	**0.992**	0.506
usaco	0.238	0.253	0.392	0.429	0.478	**0.993**	0.432
assistantbench	0.362	0.445	0.287	0.260	0.521	0.911	0.305
corebench_hard	**0.636**	**0.941**	**0.828**	**0.817**	0.805	0.972	**0.866**
onlinemind2web	0.296	0.228	0.205	0.203	0.430	0.972	0.210
scienceagentbench	0.558	0.753	0.567	0.561	**0.881**	0.982	0.610

Table 17: Kendall's tau correlation matrix of benchmarks and summary statistics for each benchmark's mean accuracy, the unweighted HAL mean score, and reliability-adjusted latent ($\widehat{\theta}$) model effect (median posterior)

	ASSISTANT	CORE-HARD	GAIA	MIND2WEB	SCICODE	SCIENCEAGENT	SWE-MINI	τ-AIRLINE	USACO	MEAN SCORE	$\widehat{\theta}$
assistantbench	1.0	0.300	**0.485**	-0.061	0.103	0.215	-0.404	0.180	0.215	0.205	0.261
corebench_hard	0.300	1.0	0.487	0.400	0.492	0.458	0.148	0.581	0.167	0.687	**0.835**
gaia	0.485	0.487	1.0	0.028	0.260	0.479	0.215	0.363	0.182	0.453	**0.606**
onlinemind2web	-0.061	0.400	0.028	1.0	0.171	0.056	0.424	0.315	-0.141	**0.581**	0.348
scicode	0.103	**0.492**	0.260	0.171	1.0	0.184	0.117	0.477	0.225	0.202	0.472
scienceagentbench	0.215	0.458	0.479	0.056	0.184	1.0	0.129	0.587	-0.085	0.392	**0.499**
swebench_verified_mini	-0.404	0.148	0.215	0.424	0.117	0.129	1.0	0.023	0.032	0.509	**0.554**
taubench_airline	0.180	0.581	0.363	0.315	0.477	0.587	0.023	1.0	0.073	0.404	**0.588**
usaco	0.215	0.167	0.182	-0.141	0.225	-0.085	0.032	0.073	1.0	**0.390**	0.338

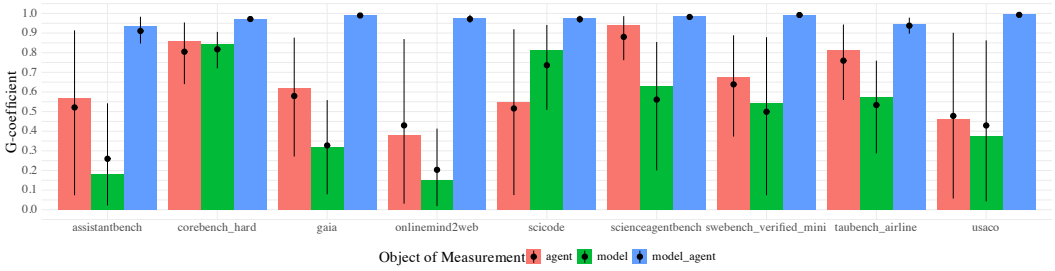


Figure 8: Reliability by Object of Measurement across all tasks in a benchmark. Estimated using variance components of Equation 2, and posteriors of Equations 3 and 4. Intervals are pseudo-standard error at 0.68 HDI, bar estimates are medians and points are means of posterior draw distributions.

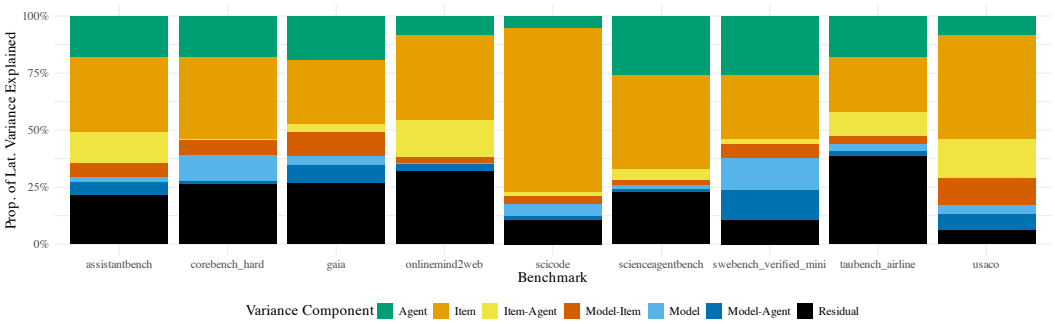


Figure 9: Proportion of Variance Explained in the posterior distributions from Eq. 2. "Item" from benchmarking literature represents "task" commonly found in agentic evaluation literature. "Agent" represents "scaffold".

the ranking obtained from the complete dataset using Spearman's rank correlation coefficient (ρ) and Kendall's rank correlation coefficient (τ). Repeating this procedure 500 times across many random subsamples yields a distribution of rank correlations for each k, quantifying how stable the model rankings are with respect to the number and selection of evaluated tasks. Task subsampling corroborates the diminishing returns. Across 500 paired subsamples, 15 tasks per benchmark retain mean Spearman agreement of 0.932 with the complete-data ranking at an estimated 82% cost reduction; 30 tasks increase agreement to 0.976 Agreement with the full ranking establishes information retention, but makes not claim to ranking validity.

Table 18: **Task subsampling shows rank information preservation.** Bootstrapped correlations of subsampled tasks and leaderboard Ranks

k	mean ρ	median ρ	ρmin	ρmax	mean τ	median τ	τmin	τmax	top@1 agreement	mean abs rank change	cost savings
5	0.790	0.801	0.584	0.917	0.634	0.640	0.458	0.766	0.640	7.221	0.94
10	0.891	0.896	0.804	0.952	0.742	0.745	0.646	0.829	0.756	5.110	0.88
15	0.932	0.936	0.870	0.966	0.799	0.802	0.720	0.858	0.834	4.052	0.82
20	0.954	0.957	0.915	0.978	0.837	0.839	0.778	0.887	0.912	3.323	0.76
25	0.967	0.969	0.940	0.982	0.863	0.865	0.811	0.901	0.960	2.822	0.69
30	0.976	0.977	0.957	0.988	0.886	0.889	0.842	0.923	0.990	2.385	0.63

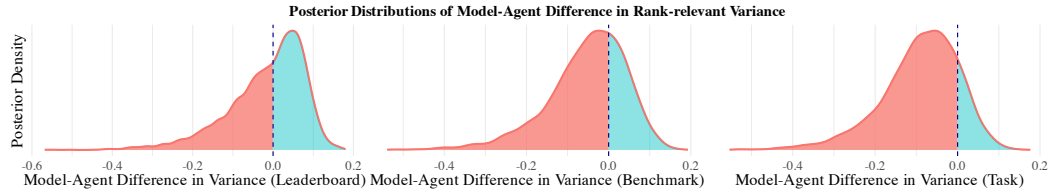


Figure 10: Posterior density of Model-Agent difference of rank-relevant variance across the leaderboard (left), across a sampled benchmark (middle), and across a sampled task (right). Positive numbers suggest LLM variance contribution is greater than that of agent scaffold.

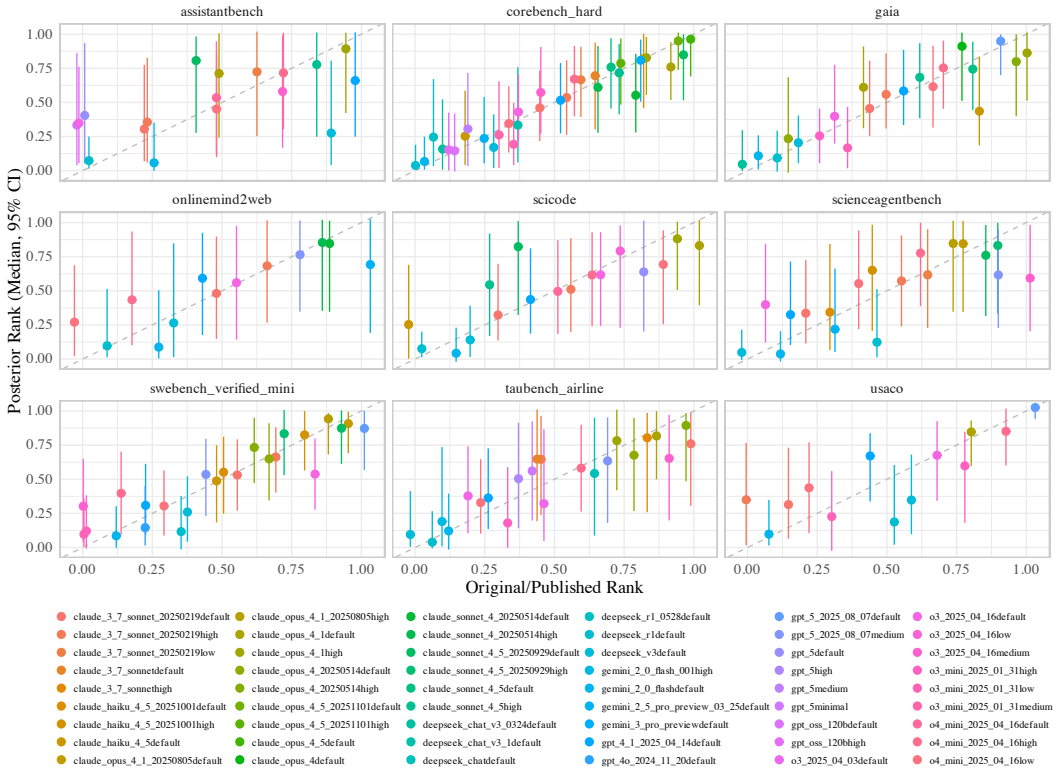


Figure 11: Published score versus posterior model percentile ranks using model-wise mean published score (x-axis) and draw-wise posterior median rankings with 95% CIs. Vertical intervals are 95% posterior rank intervals; jitter separates horizontal ties.

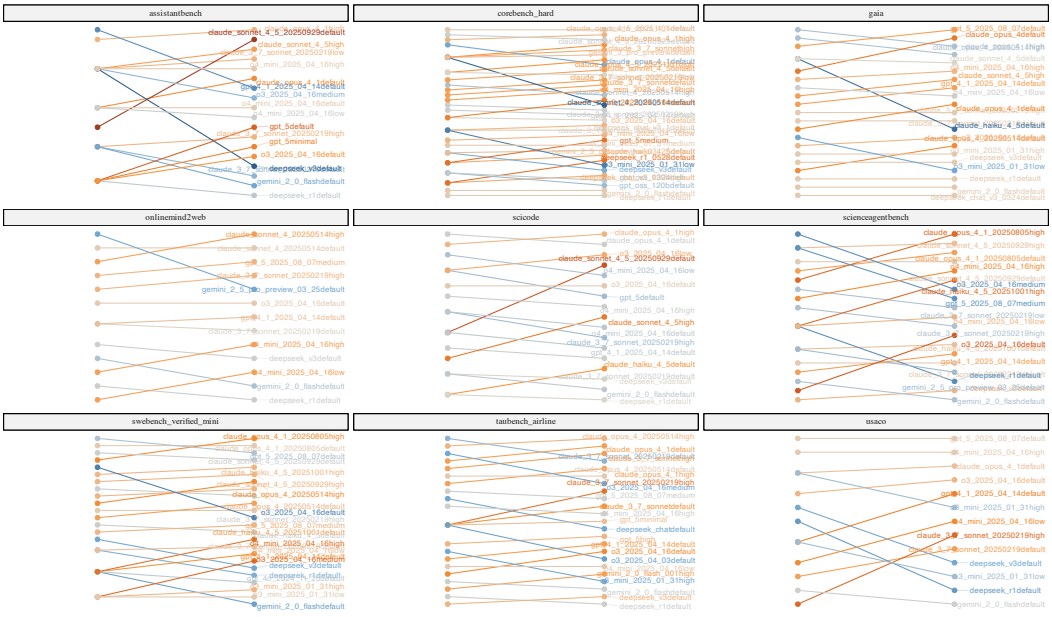


Figure 12: Rank order changes by benchmark based on comparison with median posterior rank from full variance decomposition model of Eq. 6.

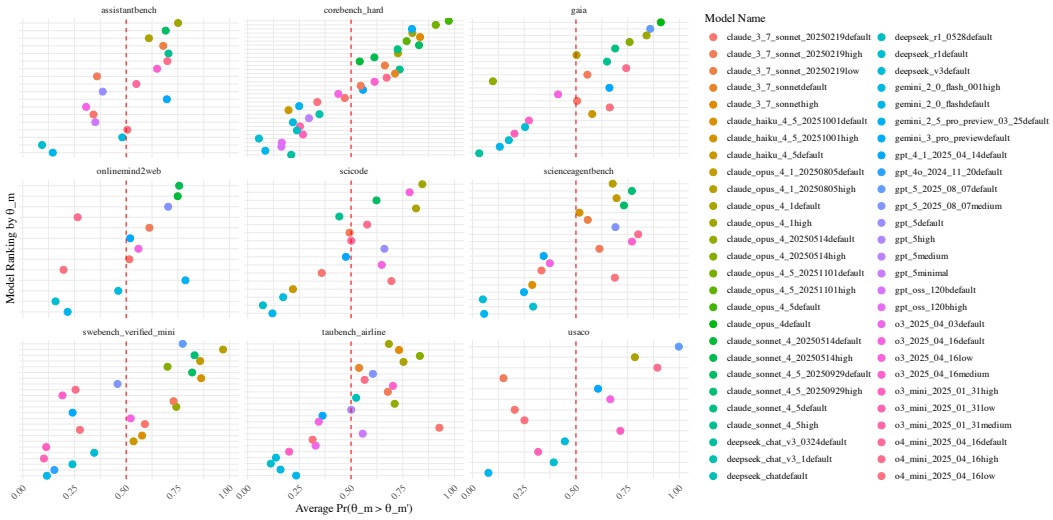


Figure 13: Cumulative posterior dominance rankings across individuals. The point estimate for each individual m represents their global dominance score, defined as the expected probability that their latent parameter $\theta_m = u_m^{(M)} + u_{mb}^{(}MB)$ is strictly greater than that of a randomly selected peer m' from the population, given the observed data y: Global Dominance$(m) = \frac{1}{M-1}\sum_{m' \neq m} \Pr(\theta_m > \theta_{m'} \mid y)$where M is the total number of individuals, and the pairwise probabilities are empirically derived via Bayesian draws as:$\Pr(\theta_m > \theta_{m'} \mid y) \approx \frac{1}{D}\sum_{d=1}^{D} \mathbb{I}\left(\theta_m^{(d)} > \theta_{m'}^{(d)}\right)$with $\mathbb{I}(\cdot)$ denoting the indicator function and D representing the total number of posterior draws. Individuals are arranged along the vertical axis in ascending order of their global dominance scores. The vertical dashed line at $\Pr = 0.5$ denotes the theoretical baseline of a perfectly average individual who exhibits structural parity relative to the rest of the cohort. Deviations toward 1.0 signify deterministic stochastic dominance over the population, whereas values approaching 0.0 indicate systematic underperformance relative to the group.



Figure 14: Pairwise Posterior Dominance, where color is Pr(Row > Column$|\ y$) across posterior draws. The point estimate for each individual m represents $\Pr(\theta_m\ >\ \theta_{m'}\ |\ y)$, defined by the expected probability of latent parameter $\theta_m = u_m^{(M)} + u_{mb}^{(}MB)$, for each benchmark b.

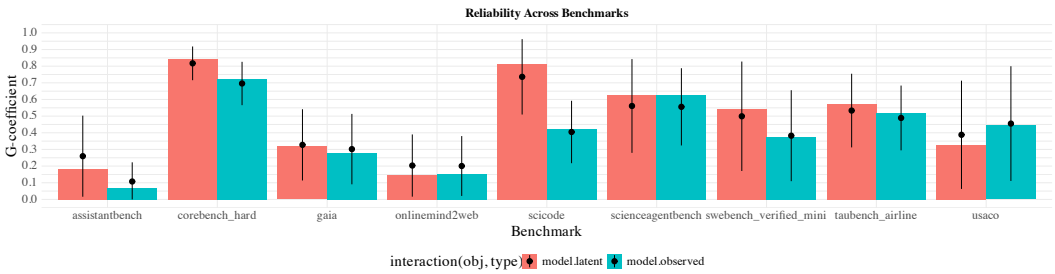


Figure 15: Rank order reliability, G-coefficient, per benchmark, for both latent and observed estimations with LLM as object of measurement. Bars are medians, points are means, and intervals are pseudo-standard error 0.68 HDI across the posterior using Eq. 3

H Method Comparison

H.1 Latent Space Estimations

Variance components in our generalized linear mixed model (GLMM) decompositions live on the *latent* (logit) scale, so ρ^2 should be interpreted as reliability of the linear predictor η (i.e., of differences in *log-odds*) rather than as reliability of raw percent-correct. This is standard in binary-response measurement: the latent scale is the scale on which additive random effects and G-theory D-study algebra apply, and it is the scale on which "true score + error" is well-defined without probability-dependent heteroskedasticity. Importantly, a given amount of latent variability implies different variability in percent-correct depending on where a model sits on the sigmoid: near $p \approx 0.5$, small changes in η translate to large changes in p, while near $p \approx 0$ or 1 they compress. For that reason, a latent-scale reliability such as $\rho^2 \approx 0.45$ *does* correspond to meaningful rank instability on the probability scale. The latent-scale ρ^2 is a mathematically coherent reliability target for binary data, and the associated posterior predictive mapping quantifies how it manifests as practically relevant leaderboard instability in percent-correct.

In addition to the interpretation benefits and alignment to measurement theoretic latent modeling, the logistic mixed effect approach also handles unbalanced data and values near zero, both common in leaderboard testing, better than the observation-level linear approach. This can be seen in Figure 15, which shows the latent and observed modeling estimated reliabilities. Rank order reliabilities are higher for the latent estimations proportionate to the quantity of low scoring models (mean score per model $\bar{y}_{mb} \lessapprox 0.1$). This is particularly true of SciCode where 100% of models have an accuracy score less than 0.1.

H.1.1 Variance shares and reliability on the latent scale

For GLMM/Bayesian GLMM, we report variance shares as proportions of total latent variance:

$$\pi_k = \frac{\sigma_k^2}{\sum_j \sigma_j^2},$$

where the sum runs over included random effects and (optionally) a latent residual term. For Bayesian fits, π_k is computed per posterior draw, yielding credible intervals. These can be found in Tables 19 and 20.

D-study scaling. When an interaction term involves a sampled facet, its contribution to the variance of an averaged score shrinks with the number of sampled levels. For example, with tasks averaged (n_i tasks), an $I \times M$ component contributes σ_{IM}^2/n_i.

H.2 DISTANCE COMPONENTS (DISCO) QUASI-RELIABILITY

The DISCO decomposition (Rizzo and Székely, 2010) generalizes classical ANOVA to arbitrary metric spaces. For K groups with n_k observations each, define the within-group dispersion as

$$\mathcal{S}_W = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i<j} \|X_{ki} - X_{kj}\|, \tag{18}$$

and the total dispersion as

$$\mathcal{S}_T = \frac{1}{N} \sum_{i<j} \|X_i - X_j\|, \tag{19}$$

where $N = \sum_k n_k$. The between-group component is $\mathcal{S}_B = \mathcal{S}_T - \mathcal{S}_W$, and the proportion attributable to the grouping factor is $\mathcal{S}_B / \mathcal{S}_T$.

For multi-facet designs, we compute DISCO sequentially for each facet, using the dispersion ratio as the analog of η^2. Because DISCO does not produce a residual term, the facet proportions do not generally sum to one when computed independently; we normalize to aid comparison with the parametric estimates.

Remark H.1. DISCO attributes substantially more dispersion to interaction terms than parametric methods. This occurs because pairwise distances capture nonlinear dependencies–such as a model performing anomalously well on a specific cluster of tasks–that additive random-effects models cannot represent. The DISCO interaction estimates should therefore be interpreted as an upper bound on the importance of non-additive effects, complementing the more conservative parametric decomposition.

DISCO decomposes dispersion using pairwise distances rather than squared deviations, improving robustness under non-normality and heavy-tailed random effects. For an object grouping g (e.g., model or model–scaffold pair), DISCO yields between-group and within-group dispersion terms (T_g, T_{within}), from which we define

$$E\rho^2 \;=\; \frac{T_g}{T_g + T_{\text{within}}}.$$

We use DISCO primarily as a methodological sensitivity analysis to validate that conclusions (dominant facets; low model-ranking reliability) are not artifacts of Gaussian random-effect assumptions.

H.3 WHY ESTIMATION METHOD CHANGES CONCLUSIONS—AND WHAT TO DO ABOUT IT

The five estimation methods are not interchangeable, and their disagreements are informative.

A notable outcome is that variance shares differ across linear mixed model (LMM), Bayesian LMM, generalized linera mixed model (GLMM), Bayesian GLMM, and DISCO, especially under sparse observations:

- **LMM** tends to allocate a large portion to residual σ^2, which can understate structured interactions when the link is misspecified for Bernoulli data.
- **GLMM** can collapse some components toward zero (boundary estimates) in unbalanced designs, yielding deceptively "clean" decompositions.
- **Bayesian LMM** allows for estimates of reliability and distributional uncertainty on observed scale, but less equipped for floor or ceiling effects or highly imbalanced data.
- **Bayesian GLMM** exposes skewness and large uncertainty, which is appropriate when the design under-identifies components.
- **DISCO** can attribute more to interaction-like dispersion without parametric assumptions, often aligning with the intuition that agentic pipelines create nonlinear, heteroskedastic effects.

Practical guidance.

1. Use **Bayesian** or **nonparametric** methods to communicate uncertainty when data are sparse; prefer point-estimate GLMM/LMM only when coverage is dense.
2. Triangulate: when all four methods agree on the *ranking* of dominant facets (e.g., task vs interactions vs residual), recommendations are robust; when they disagree, the correct conclusion is that the benchmark is under-instrumented for that inference.

Linear vs. generalized models. The LME estimates generally attribute the largest share of variance to the residual (with a borderline exception of USACO), with correspondingly compressed named components. The GLME and Bayesian estimates, by contrast, allocate substantially more variance to specific tasks and models. This divergence is expected: the Gaussian identity-link model treats binary $\{0, 1\}$ responses as continuous, and its residual absorbs the Bernoulli variance floor ($p(1-p)$) that the logit-link models separate out via the distributional assumption. The practical implication is that *linear G-theory estimates applied to pass/fail benchmarks systematically underestimate the proportion of variance attributable to named facets and overestimate residual noise*, leading to artificially compressed—but not necessarily more conservative—reliability estimates.

Bayesian vs. frequentist GLME. The Bayesian and GLME estimates agree in broad strokes but diverge in two instructive ways. First, the Bayesian posteriors for agent variance are markedly right-skewed, with posterior means 2–5$\times$ larger than posterior medians (e.g., SWE-bench agent mean $= 0.27$, median $= 0.19$). The GLME point estimate, which approximates the posterior mode, misses this tail mass and underestimates the expected contribution of scaffold variance. Second, the credible intervals from the Bayesian analysis expose the *decision-relevant uncertainty*: for several benchmarks, the 95% HDI for model variance includes zero, meaning the data are consistent with no true model differentiation at all. This uncertainty is invisible in frequentist analyses.

Nonparametric DISCO. The DISCO estimates depart most dramatically from the parametric methods on the interaction terms, difference likely due to heterogeneous cluster imbalances (and hence our preference for Bayesian estimates). DISCO attributes 40–57% of total dispersion to task$\times$model interactions across benchmarks, compared to 1–11% from the Bayesian estimates. This divergence likely reflects two mechanisms: (i) DISCO's sensitivity to nonlinear dependencies that the additive random-effects models cannot capture, and (ii) the absence of a separate residual term in DISCO, which forces unexplained variation into the named interactions. The practical upshot is that DISCO serves as a useful *upper bound* on interaction effects and a reminder that the additive decomposition assumed by mixed-effects models may understate the complexity of the task$\times$model relationship.

H.4 Variance decomposition explains the reliability ceiling and generalizability

Figure 16 reports posterior variance shares from the leaderboard model. Persistent model variation is smaller than the combined variation associated with scaffolds, benchmark-conditioned model performance, and item-conditioned interactions. Most observed item outcomes therefore contain more condition-specific variation than globally transferable model signal.

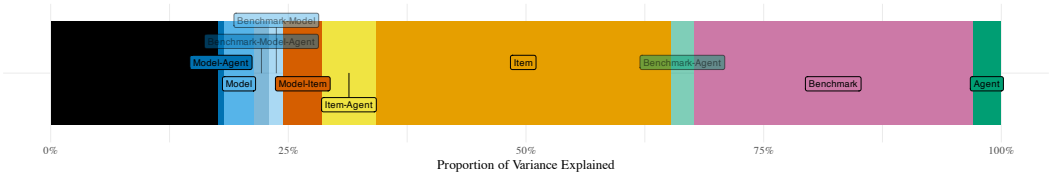


Figure 16: Full variance decomposition of Eq. 6 as percentages of total variance. Variance decompositions for individual benchmarks are in Appendix G. Black represents residual variance.

This does not imply that model capability is unimportant. Reliability is population-dependent: contemporary frontier models are often similar, whereas tasks and scaffolds are heterogeneous. A benchmark could have ranked a historically broader model population reliably and yet fail to distinguish the current frontier. Likewise, a future benchmark may become uninformative as models saturate it. Reliability must consequently be re-estimated as the evaluated population and agent ecosystem change.

For strong generalization to new tasks, we would want to see larger main effects for models and scaffolds, making decisions clearer. However, the reality is that the second-order task-level interactions, model–task and scaffold–task, represent larger variation components than their respective main effects. This quantifies that, instead of providing clear architecture choices, some models

or scaffolds are better at certain tasks than others, implying that a developer should, for a given task, choose a specific model and scaffold for each new task, rather than for the group of tasks represented by a benchmark or leaderboard (because of a smaller benchmark–model component). This is not ideal for developers, as it means they would need to know about the demands of each task and the appropriateness of architecture selections for it. This partly explains why there is such low measurement reliability for fewer numbers of items in Figures 1 and 3: a different model scaffold might be optimal for each task item.

H.5 Absolute Reliability on Linear Scale

Generalizability theory also permits estimation of reliability of numeric scores. Absolute reliability, as it is known, is always less than or equal to the corresponding estimate of relative reliability. But because scores are on the observed scale, all of the analyses in this section are based on the Bayesian linear mixed model (§ C.2). They are illustrative and serve as a framework for practitioners looking to measure reliability of assigned numeric scores.

Reliable *rankings* do not imply reliable *scores*. Relative reliability ($E\rho^2$) asks whether the ranking of models stays similar across evaluation conditions, while absolute reliability (Φ) asks whether their reported scores stay similar. Figure 18 shows that this distinction matters substantially in practice: absolute reliability is consistently lower across the nine benchmarks, and for several benchmarks remains near zero even as ranking reliability increases with additional tasks. An evaluation can therefore support a stable ranking without supporting equally strong claims about a model's absolute score or whether it exceeds a fixed threshold.

While not recommended for the HAL dataset and benchmarks, there may be some agentic benchmarks where absolute scores are needed for decision-making. Absolute reliability is most interpretably performed on the observed scale, rather than the latent scale used throughout the main body of this study. For the HAL dataset, there is simply not enough signal relative to noise in the benchmarks for any reliable absolute scoring. Nevertheless, the concept of absolute reliability may be important for some agentic measures.

H.5.1 Absolute reliability

Rank-based reliability measures whether an evaluation preserves the ordering of models or systems, but it does not establish whether their score levels are reproducible. Absolute reliability is important when scores are interpreted directly—for example, when assessing whether a model exceeds a deployment threshold, meets a minimum capability requirement, or achieves a specified performance target. In these settings, a task set or scaffold that shifts every model's score equally can change the decision even without changing the ranking. The absolute reliability, or dependability coefficient, Φ, therefore includes these common shifts in its error variance.

Treating tasks and scaffolds as random facets, the absolute reliability of a model score under an equal-allocation design is

$$\Phi_{M(b)}(n_i, n_a) = \frac{\sigma^2_M}{\sigma^2_M + (\sigma^2_I + \sigma^2_{IM})/n_i + (\sigma^2_A + \sigma^2_{MA})/n_a + (\sigma^2_{IA} + \sigma^2_{IMA,e})/(n_i n_a)}. \quad (20)$$

Relative to model-ranking reliability, the denominator additionally includes task main effects, scaffold main effects, and task–scaffold interactions. These components affect the absolute score of a model averaged over sampled tasks and scaffolds, even when they do not affect its position relative to other models. Increasing the number of tasks reduces task-related error, whereas increasing the number of scaffolds reduces scaffold-related error.

When the object of measurement is the complete *model–scaffold system*, scaffold differences are part of the system's universe score rather than measurement error. Absolute reliability is then

$$\Phi_{MA(b)}(n_i) = \frac{\sigma^2_M + \sigma^2_A + \sigma^2_{MA}}{\sigma^2_M + \sigma^2_A + \sigma^2_{MA} + (\sigma^2_I + \sigma^2_{IM} + \sigma^2_{IA} + \sigma^2_{IMA,e})/n_i}. \quad (21)$$

Here, task main effects enter the denominator because sampling an easier or harder task set can shift a system's score across an absolute decision threshold. In contrast, scaffold main effects and model–scaffold interactions remain signal because they distinguish the systems being evaluated.

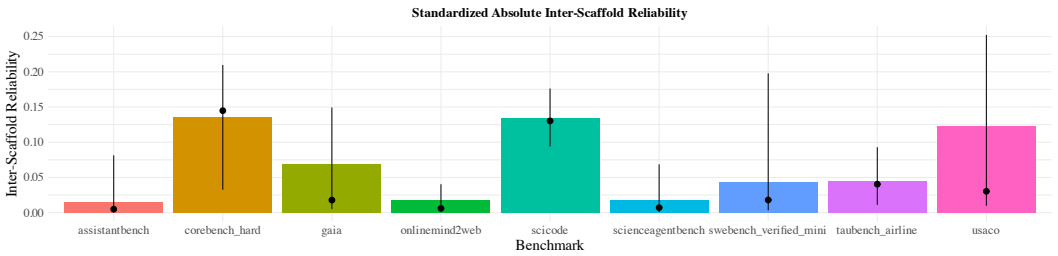


Figure 17: Standardized Absolute Inter-scaffold Reliability. Standardized inter-scaffold reliability looks at the reliability for a single fixed task, rather than the composite reliability across all tasks in a benchmark.

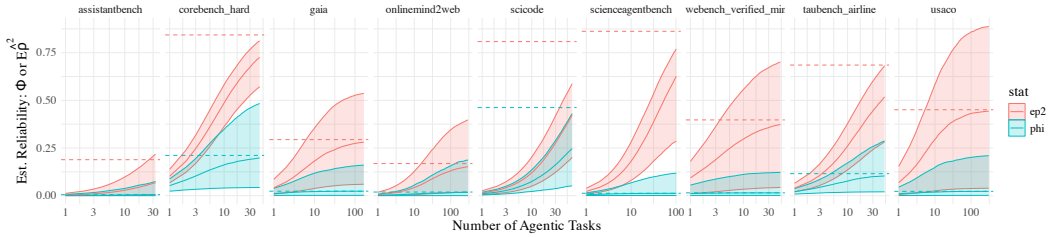


Figure 18: Absolute vs Relative Reliability on Observed Scale

For either object of measurement, absolute reliability is no greater than the corresponding ranking reliability:

$$\Phi_{M(b)}(n_i, n_a) \leq E\rho^2_{M(b)}(n_i, n_a), \qquad \Phi_{MA(b)}(n_i) \leq E\rho^2_{MA(b)}(n_i). \tag{22}$$

A large gap indicates that rankings may be reproducible even though reported score levels are sensitive to the sampled evaluation conditions. Likewise, high inter-scaffold ranking reliability does not guarantee absolute agreement: two scaffolds can preserve model ordering while producing systematically different scores.

H.6 Full Estimation Tables for Proportion of Latent Variation Explained

Full variance-proportion tables for all models and all five estimation methods are provided in the following tables including Bayesian posterior summaries and also LME, GLME, and DISCO estimates to enable cross-method comparison.

Table 19: Proportion of variation explained per benchmark. For comparability across methods, Bayesian estimates represent posterior means (rather than medians found in most of the study) to preserve proportional relationships.

Benchmark	Facet	LMM	Bayes LMM	GLMM	Bayes GLMM	DISCO
scicode	item	0.235	0.232	0.766	0.718	0.209
scicode	agent	0.001	0.041	0.004	0.054	0.001
scicode	model	0.010	0.011	0.053	0.057	0.012
scicode	item_agent	0.037	0.036	0.012	0.016	0.252
scicode	item_model	0.130	0.123	0.012	0.035	0.509
scicode	model_agent	0.002	0.004	0.003	0.015	0.017
scicode	sigma	0.584	0.553	0.150	0.105	NA
gaia	item	0.210	0.113	0.342	0.280	0.169
gaia	agent	0.041	0.471	0.029	0.194	0.017
gaia	model	0.030	0.020	0.040	0.042	0.046
gaia	item_agent	0.036	0.020	0.035	0.036	0.209
gaia	item_model	0.097	0.051	0.041	0.102	0.481
gaia	model_agent	0.059	0.039	0.083	0.079	0.077
gaia	sigma	0.527	0.286	0.431	0.267	NA
taubench_airline	item	0.166	0.140	0.277	0.244	0.161
taubench_airline	agent	0.043	0.237	0.044	0.180	0.021
taubench_airline	model	0.027	0.024	0.039	0.035	0.025
taubench_airline	item_agent	0.086	0.070	0.107	0.101	0.250
taubench_airline	item_model	0.035	0.024	0.000	0.033	0.484
taubench_airline	model_agent	0.012	0.012	0.016	0.020	0.059
taubench_airline	sigma	0.631	0.493	0.518	0.386	NA
swebench_verified_mini	item	0.153	0.070	0.349	0.277	0.092
swebench_verified_mini	agent	0.246	0.645	0.103	0.260	0.073
swebench_verified_mini	model	0.056	0.027	0.272	0.142	0.119
swebench_verified_mini	item_agent	0.068	0.034	0.024	0.026	0.190
swebench_verified_mini	item_model	0.089	0.039	0.009	0.059	0.399
swebench_verified_mini	model_agent	0.057	0.036	0.042	0.130	0.126
swebench_verified_mini	sigma	0.331	0.149	0.202	0.106	NA
usaco	item	0.257	0.141	0.551	0.450	0.239
usaco	agent	0.072	0.469	0.000	0.090	0.003
usaco	model	0.062	0.022	0.000	0.049	0.027
usaco	item_agent	0.200	0.114	0.130	0.172	0.262
usaco	item_model	0.184	0.103	0.000	0.116	0.440
usaco	model_agent	0.000	0.027	0.103	0.066	0.030
usaco	sigma	0.225	0.124	0.216	0.057	NA
assistantbench	item	0.116	0.074	0.437	0.331	0.156
assistantbench	agent	0.000	0.377	0.000	0.181	0.002
assistantbench	model	0.000	0.003	0.000	0.024	0.017
assistantbench	item_agent	0.083	0.057	0.107	0.137	0.219
assistantbench	item_model	0.047	0.022	0.000	0.057	0.574
assistantbench	model_agent	0.013	0.009	0.069	0.058	0.032
assistantbench	sigma	0.741	0.458	0.386	0.213	NA
corebench_hard	item	0.228	0.178	0.415	0.362	0.153
corebench_hard	agent	0.063	0.286	0.047	0.180	0.027
corebench_hard	model	0.071	0.056	0.134	0.113	0.074
corebench_hard	item_agent	0.010	0.009	0.000	0.005	0.200
corebench_hard	item_model	0.099	0.074	0.029	0.063	0.462
corebench_hard	model_agent	0.012	0.012	0.011	0.016	0.083
corebench_hard	sigma	0.517	0.385	0.365	0.262	NA
onlinemind2web	item	0.283	0.205	0.414	0.372	0.239
onlinemind2web	agent	0.000	0.272	0.000	0.085	0.000
onlinemind2web	model	0.002	0.004	0.003	0.009	0.007
onlinemind2web	item_agent	0.133	0.097	0.164	0.163	0.296
onlinemind2web	item_model	0.022	0.014	0.000	0.023	0.446
onlinemind2web	model_agent	0.016	0.013	0.027	0.029	0.012
onlinemind2web	sigma	0.545	0.395	0.393	0.319	NA
scienceagentbench	item	0.246	0.129	0.532	0.413	0.208
scienceagentbench	agent	0.042	0.505	0.048	0.258	0.011
scienceagentbench	model	0.016	0.008	0.026	0.020	0.016
scienceagentbench	item_agent	0.086	0.045	0.049	0.046	0.249
scienceagentbench	item_model	0.034	0.017	0.000	0.021	0.493
scienceagentbench	model_agent	0.000	0.003	0.008	0.014	0.023
scienceagentbench	sigma	0.577	0.294	0.338	0.227	NA

Table 20: Proportion of variation explained full leaderboard. For comparability across methods, Bayesian estimates represent posterior means (rather than medians found in most of the study) to preserve proportional relationships.

Facet	LMM	Bayes LMM	GLMM	Bayes GLMM	DISCO
agent	0.016	0.039	0.020	0.040	0.014
benchmark	0.075	0.108	0.259	0.299	0.023
benchmark_agent	0.028	0.036	0.026	0.031	0.030
benchmark_item	0.252	0.234	0.328	0.297	0.117
benchmark_item_agent	0.070	0.065	0.052	0.055	0.140
benchmark_item_model	0.047	0.043	0.004	0.040	0.240
benchmark_model	0.010	0.008	0.020	0.015	0.040
benchmark_model_agent	0.013	0.014	0.014	0.016	0.046
model	0.026	0.025	0.033	0.032	0.015
model_agent	0.007	0.006	0.009	0.008	0.035
benchmark_task_model_agent + sigma	0.456	0.422	0.236	0.168	0.302