How Reliable Are Agent Leaderboards? A Variance-Decomposition Analysis

ANONYMOUS AUTHOR(S)

CCS Concepts: • **Computing methodologies** → **Machine learning**; Natural language processing; • **General and reference** → **Evaluation**; *Measurement*; • **Mathematics of computing** → *Multilevel and mixed-effects models.*

Additional Key Words and Phrases: reliability, evaluations, generalizability theory, rankings, leaderboards

ACM Reference Format:

Anonymous Author(s). 2026. How Reliable Are Agent Leaderboards? A Variance-Decomposition Analysis. In *Proceedings of ACM CAIS 2026: RLEval Workshop.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Leaderboards on benchmarks like SWE-bench [12], GAIA [17], and τ-bench [30] are referenced in technical reports, system cards, and policy documents [1, 18], and a model[1] that ranks first is treated as the state of the art for the capability the benchmark purports to measure. Yet the reliability, i.e., how stable these rankings are, has not been established. Reliability is not a property of a benchmark in the abstract: it depends on *what the benchmark is being used to rank*. "Model X achieves 47% on SWE-bench" under a single scaffold can be a claim about the base model or about the (model, scaffold) system, and current practice often does not state which is being evaluated. Applying generalizability theory [3, 6] to seven agent benchmarks on the Holistic Agent Leaderboard (HAL) [14], we decompose score variance into facets corresponding to model, item, scaffold, and their interactions, and compute reliability for two natural targets: ranking models and ranking model–scaffold pairs. We find that:

- **Model and model–scaffold reliability diverge on the same data.** On AssistantBench, GAIA, and USACO, model–scaffold rankings are reliable while model rankings are not; on CoreBench-Hard, the reverse holds. Current reporting does not distinguish these objects.
- **Item scaling alone cannot raise the model-ranking ceiling.** The model main effect—the signal a leaderboard is meant to surface—accounts for only ∼3% of total score variance. Within a single benchmark, reliability asymptotes near $E\hat{\rho}^2 \approx 0.52$; broadening across seven HAL benchmarks raises the projected ceiling to ≈ 0.88. The dominant variance components attenuate with task variety and scaffold count, not item count. Reliable model rankings therefore require evaluation breadth, not necessarily more items.
- **Published leaderboards are often statistically indistinguishable.** Median inter-draw Spearman correlations of model capability estimates range from 0.16 to 0.47, and on five of seven benchmarks no pair of model ranks are mutually distinguishable at $\alpha = 0.1$. Re-ranking reorders the leaderboards substantially: the model published as best on τ-bench Airline drops 16 ranks, and the SWE-bench Verified Mini leader drops to rank 12.

[1]Throughout, "model" refers to a base language model paired with a specific reasoning-effort configuration; e.g., `gpt-5-high` and `gpt-5-minimal` are treated as distinct models.

Manuscript submitted to ACM

We close with four principles for the design, reporting, and interpretation of agent evaluations.

2 Generalizability theory for agent benchmarks

G-theory decomposes observed-score variability into variance components attributable to facets and their interactions, and defines reliability as the fraction of that variability attributable to the object of measurement rather than to sampling of nuisance facets [3, 6]. We specify the commitments G-theory requires, then state the two decompositions we use in the main body. Two intermediate decompositions and full estimation details are in Appendix I, along with a more detailed methodology description. A score is indexed by benchmark b, item i, model m, and agent scaffold a, with $y_{b,i,m,a} \in \{0, 1\}$. We treat all four facets as random. Items are sampled from each benchmark's published item set. Scaffolds are sampled from the 2–3 scaffolds reported in current evaluation practice per benchmark; this is a field-wide pattern, not a feature of HAL, and we pool across benchmarks below to recover power. Models are sampled from frontier models evaluated on HAL; this universe is clustered (shared architectures and training pipelines), which depresses σ^2_M as a feature of the current model landscape.

We compute reliability for two objects of measurement: the model m (with scaffold treated as nuisance) and the model–scaffold pair (m, a) (treated as the deployable unit). Because leaderboard use is a relative-decision problem, we report relative reliability $E\rho^2$ throughout, in which nuisance-facet main effects shared across objects drop out[2] and only object-by-facet interactions and residual terms enter the error:

$$E\rho^2 = \frac{\sigma^2_o}{\sigma^2_o + \sigma^2_{\text{rel-error}}(n_i, n_a, n_b)}, \tag{1}$$

where $\sigma^2_{\text{rel-error}}(\cdot)$ is the decision-study (D-study) projection describing how error shrinks as sampled items, scaffolds, or benchmarks grow. Full per-decomposition expressions are in Appendix I.8.

2.1 Two decomposition models

Let $\eta_{b,i,m,a}$ denote the logit-scale linear predictor, with $y_{b,i,m,a} \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_{b,i,m,a}))$. We present two decompositions in the main body; two intermediate per-benchmark decompositions (one with (m, a) as the object, one separating model, agent, and their interactions) are in Appendix I. *Decomposition 1 (naïve per-benchmark)* mirrors the implicit assumption of current leaderboard practice: scores vary by model and item, with all other variability absorbed into residual. For each benchmark b, $\eta_{i,m} = \beta_0 + u^{(I)}_i + u^{(M)}_m$. We use this as a strawman to quantify the optimism induced by ignoring scaffold variance and model–scaffold interactions. *Decomposition 2 (pooled ecosystem)* pools all benchmarks and includes benchmark-level facets and cross-benchmark interactions, anchored across benchmarks by the HAL-generalist scaffold:

$$\begin{aligned}\eta_{b,i,m,a} = \beta_0 &+ u^{(B)}_b + u^{(BI)}_{b:i} + u^{(M)}_m + u^{(A)}_a + u^{(MA)}_{m:a} \\ &+ u^{(BM)}_{b:m} + u^{(BA)}_{b:a} + u^{(BMA)}_{b:m:a} + u^{(BIM)}_{b:i:m} + u^{(BIA)}_{b:i:a} + u^{(BIMA)}_{b:i:m:a}.\end{aligned} \tag{2}$$

All effects are mean-zero Gaussian. This is the workhorse model for our central question: even if we scale item counts, is model-ranking reliability bounded by cross-benchmark and scaffold variability? Closed-form D-study projection for model-ranking reliability under this decomposition is

$$E\rho^2_M(n_b, n_i, n_a) = \frac{\sigma^2_M}{\sigma^2_M + \frac{\sigma^2_{BM}}{n_b} + \frac{\sigma^2_{MA}}{n_a} + \frac{\sigma^2_{BMA}}{n_b n_a} + \frac{\sigma^2_{BIM}}{n_b n_i} + \frac{\sigma^2_{BIMA,e}}{n_b n_i n_a}}, \tag{3}$$

[2]The framework extends to absolute reliability for evaluating scores in isolation by adding these drop-outs.

which makes explicit that item-indexed components attenuate with n_i but model-by-benchmark and model-by-scaffold components do not. We analyze agent rollouts from seven benchmarks distributed via HAL [14]: AssistantBench [31], CoreBench Hard [26], GAIA [17], SciCode [28], SWE-bench Verified Mini [12], τ-bench Airline [30], and USACO [25], covering web navigation, scientific reproducibility, multi-step assistance, scientific programming, software engineering, customer-service interaction, and competitive programming respectively (full descriptions in Appendix G). Our analysis covers 20, 144 total agent rollouts.

3 Results

3.1 Reliability depends on the unit of comparison

A practitioner may ask *which model performs best on benchmark X?* or *which model–scaffold pair performs best on benchmark X?*, and these can yield different answers on the same dataset. Figure 1 shows posterior reliability estimates for each benchmark under both objects of measurement, computed from a per-benchmark decomposition that separates model, scaffold, and their interactions (Appendix I).

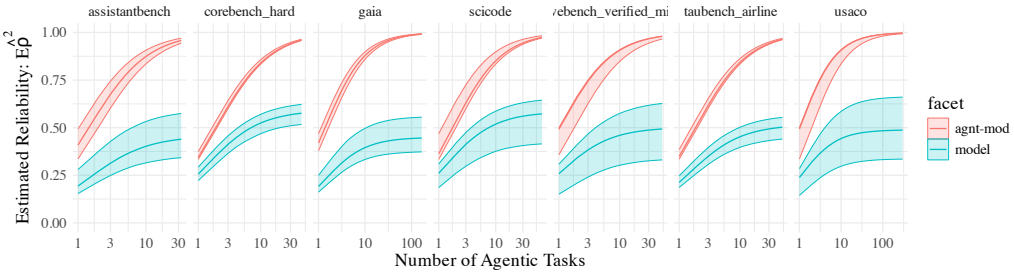

Fig. 1. Estimated reliability by number of items, comparing model and model–scaffold as objects of measurement. D-study estimates sampled from the posterior of the per-benchmark decomposition. Credible intervals are 95% HDI; effect estimates are medians. Variance components aggregated after inverse-logit transform; reliabilities represent the observation space.

Reliability varies substantially across benchmarks, and which object yields higher reliability is itself benchmark-specific. AssistantBench, GAIA, and USACO show pair-unit reliability approaching 1.0 while model-unit reliability remains below 0.5: inferences about model performance alone are weak on these benchmarks, while inferences about model–scaffold performance can be strong. CoreBench-Hard shows the opposite pattern. There is no single reliability number that characterizes an agent benchmark; the answer depends on what the benchmark is being used to rank, and current reporting practice does not make that choice explicit.

3.2 Item scaling saturates; benchmark breadth raises the ceiling

The reliability bottleneck is not the number of items but the diversity these tasks span. Figure 3 shows that when all additional tasks are drawn from a single benchmark, projected model-ranking reliability plateaus around $E\hat{\rho}^2 \approx 0.52$. This is an item-only ceiling: variance from model–benchmark and model–scaffold interactions persists even as item-level noise vanishes (Eq. 3). When the same budget is spread across the seven HAL benchmarks, benchmark-conditioned model heterogeneity is averaged down and the projected ceiling rises to ≈ 0.88. The ceiling is design-conditional, not universal, i.e., at $n_b = 1$ the benchmark–model variance does not average out; at $n_b = 7$ it is divided by seven. Figure 4 shows why the single-benchmark plateau is so low. The model main effect accounts for ∼3% of total variance under

the pooled decomposition (2–14% per-benchmark; Appendix J), smaller than item-conditioned and scaffold-related components. On a single benchmark, score differences between models are dominated by the item and scaffold draw, and a single-benchmark leader position is not statistically separable from the field (§3.3). The corollary for design is that reliable model rankings are achieved by broadening across independently informative, diverse tasks and scaffolds, not by blindly adding more items.

3.3 Consequences for current agent leaderboards

We translate these variance components into rank uncertainty on the HAL leaderboards. Median pairwise inter-draw Spearman correlations of BLUP rankings across the posterior are low, ranging from 0.16 (AssistantBench) to 0.47 (SWE-bench Verified Mini); see Appendix A. Published orderings reflect a particular draw of items and scaffolds, and a different draw could plausibly invert them. Across all seven benchmarks, only GAIA and SWE-bench Verified Mini produce any pair of model capability estimates that are mutually distinguishable at $\alpha = 0.1$. Figure 2 compares published rankings (the naïve decomposition implicit in current practice) with rankings under the pooled decomposition. The model published as best on SWE-bench Verified Mini falls to rank 12; on τ-bench Airline the largest single rank shift is 16 positions. The naïve ranking conflates model capability with model–scaffold compatibility: a model paired with a well-fitting scaffold is credited for the fit. The pooled decomposition separates the two, and the reordering shows which models were over- or under-credited by the scaffolds they were tested with.

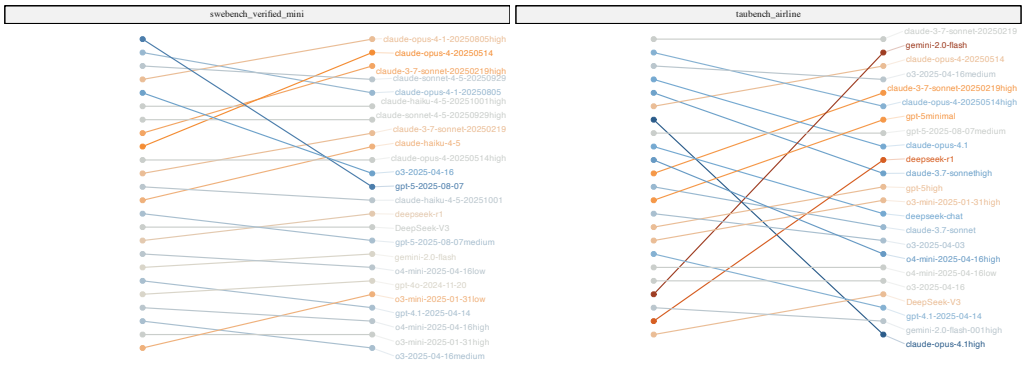

Fig. 2. Rankings on τ-bench Airline and SWE-bench Verified Mini under current practice (left) and under BLUPs from the pooled decomposition (right). Remaining benchmarks in Fig. 10.

Four principles follow from our findings[3]: (1) **Declare the object of measurement**: scores should specify what they rank (model, model–scaffold pair, or other unit), since this choice determines signal versus noise and yields divergent reliability on the same data (§3.1). (2) **Attribute improvements to the factor that varied**: releases that simultaneously change base model, scaffold, prompting, and inference budget cannot be decomposed post hoc, and aggregate gains can reverse once contributions are separated (§3.3). (3) **Report reliability with uncertainty**: a point estimate, credible interval, declared object, and estimation method are the minimum reportable set. (4) **Diversify contexts, not just task counts**: because σ^2_{BM}, σ^2_{BMA}, and σ^2_{MA} do not attenuate with item count, budget should be allocated across independently informative, diverse tasks and multiple scaffolds, with D-study curves over n_i, n_a, and n_b reported (§3.2).

[3]Full principles in Appendix C.

References

[1] Anthropic. 2026. *System Card: Claude Opus 4.7.* System Card. Anthropic.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* 67, 1 (2015). doi:10.18637/jss.v067.i01

[3] Robert L. Brennan. 2001. *Generalizability Theory.* Springer, New York, NY. doi:10.1007/978-1-4757-3456-0

[4] Robert L Brennan. 2003. *Coefficients and Indices in Generalizability Theory.* Technical Report 1. Center for Advanced Studies in Measurement and Assessment.

[5] Paul-Christian Bürkner. 2021. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software* 100 (Nov. 2021), 1–54. doi:10.18637/jss.v100.i05

[6] Lee J. Cronbach, Goldine Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. *The Dependability of behavioral measurements: theory of generalizability for scores and profiles.* Wiley, New York.

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

[8] Lee J. Cronbach and Richard J. Shavelson. 2004. My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement* 64, 3 (June 2004), 391–418. doi:10.1177/0013164404266386

[9] Michael Hardy. 2025. Measuring Teaching with LLMs. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish (Eds.). National Council on Measurement in Education (NCME), Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States, 367–384. https://aclanthology.org/2025.aimecon-main.40/

[10] Michael Hardy and Yunsung Kim. 2026. Knowledge without Wisdom: Measuring Misalignment between LLMs and Intended Impact. doi:10.48550/arXiv.2603.00883 arXiv:2603.00883 [cs].

[11] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. 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).* 6864–6890.

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

[13] 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. 2025. Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation. doi:10.48550/arXiv.2510.11977 arXiv:2510.11977 [cs].

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

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

[16] Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying Variance in Evaluation Benchmarks. doi:10.48550/arXiv.2406.10229 arXiv:2406.10229.

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

[18] OpenAI. 2026. *GPT-5.5 System Card.* System Card. OpenAI.

[19] Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. 2026. Towards a Science of AI Agent Reliability. doi:10.48550/arXiv.2602.16666 arXiv:2602.16666 [cs].

[20] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking Prompt Sensitivity in Large Language Models. *ArXiv* abs/2502.06065 (2025). https://api.semanticscholar.org/CorpusID:276249536

[21] William Revelle and David M. Condon. 2019. Reliability from to : A tutorial. *Psychological Assessment* 31, 12 (2019), 1395–1411. doi:10.1037/pas0000754 Place: US.

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

[23] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. doi:10.48550/arXiv.2505.10573 arXiv:2505.10573 [cs].

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

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

[26] Zachary S Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. 2024. Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363* (2024).

[27] Gábor J. Székely and Maria L. Rizzo. 2017. The Energy of Data. *Annual Review of Statistics and Its Application* 4, 1 (March 2017), 447–479. doi:10.1146/annurev-statistics-060116-054026

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

[29] Ruipeng Wang, Yuxin Chen, Yukai Wang, Chang Wu, Junfeng Fang, Xiaodong Cai, Qi Gu, Hui Su, An Zhang, Xiang Wang, et al. 2026. AgentNoiseBench: Benchmarking Robustness of Tool-Using LLM Agents Under Noisy Condition. *arXiv preprint arXiv:2602.11348* (2026).

[30] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045* (2024).

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

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

Appendix

A Additional Figures

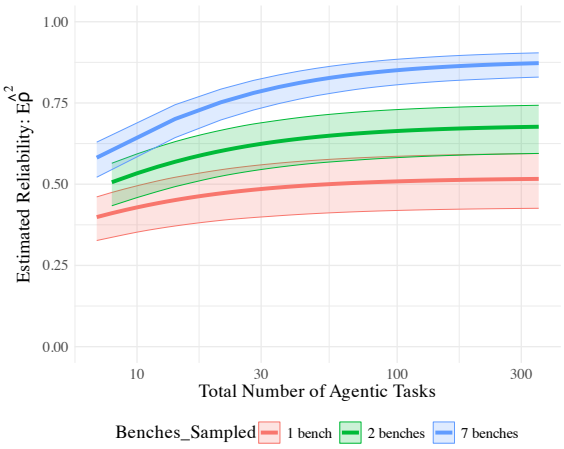

Fig. 3. Projected model-ranking reliability under the pooled decomposition (Eq. 6), holding total item budget fixed across three sampling regimes: 350 tasks from one benchmark, 175 from each of two, or 50 from each of seven.

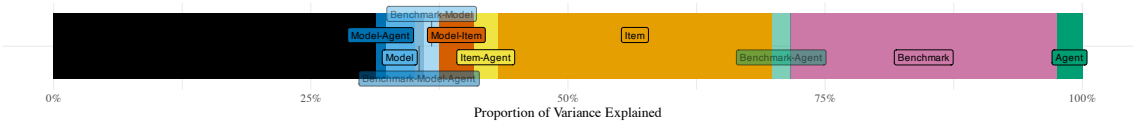

Fig. 4. Variance share of each component under the pooled decomposition (Eq. 6), computed via Eq. 12. Per-benchmark decompositions in Appendix J.

Table 1. Median pairwise inter-draw Spearman correlation of BLUP rankings across the posterior. Low values indicate that the published rank order is not stable to resampling of items and scaffolds.

AssistantBench	CoreBench-Hard	GAIA	SciCode	SWE-bench V. Mini	τ-bench Airline	USACO
0.164	0.219	0.260	0.265	0.470	0.186	0.374

B Related Work

Generalizability theory and measurement reliability. Generalizability theory (G-theory) is a measurement-theoretic framework for decomposing observed score variance into contributions from multiple facets and their interactions [3, 6, 8, 21, 24]. G-theory can be thought of as an extension of ANOVA: where ANOVA estimates how much variance is attributable to each factor, G-theory goes further by using those variance components to project reliability under hypothetical evaluation designs. Recent work has applied variance-decomposition methods to AI benchmarks more broadly, showing that design choices can induce substantial variability in rankings and conclusions [9, 10, 16].

Agent benchmarks and their reliability. A growing ecosystem of benchmarks agentic benchmark, software engineering [12], web navigation [11, 32], general-purpose assistance [17], and customer-service [30], use a leaderboard to indicate their best performing agent. These leaderboards rank using success rate alone, making it difficult to infer possible inconsistencies in the agent's performance. Several lines of work have introduced reliability metrics that surface sources of uncertainty in agent's performance. Rabanser et al. [19] and Yao et al. [30] measure consistency by rerunning task evaluations. Others test robustness under varies conditions, prompt perturbations [19, 20], tool variation [29], and environmental errors injected during rollout [15, 19]. This prior work asks whether a single agent performs consistently under reasonable variation in its task. We ask a different question: given an experimental design, how reliable are the comparative conclusions drawn from it? The answer depends not only on the agent under test but on the set of agents it is being compared against. To our knowledge, no prior work has studied the reliability of inferences drawn from agentic experiments.

C Principles for Reliable Agent Evaluation

Four principles follow from the findings in Section 3, addressed to benchmark designers, reporters, and consumers of agent evaluations.

Declare the object of measurement. Every score should declare what it ranks: the model, the model–scaffold pair, or some other unit. The choice determines which components are signal (σ^2_M vs. σ^2_{MA} absorbed into signal) and which are error, and yields divergent reliability on the same data (Section 3.1). For example, "47% on SWE-bench" under a single scaffold is a result about the (model, scaffold) system, not the model.

Attribute improvements to the factor that varied. Model releases often change base model, scaffold, tools, prompting, and inference budget at once. Aggregate gains cannot be decomposed post hoc: a model-level claim requires multiple scaffolds, a scaffold-level claim multiple base models. Section 3.3 showed the model published as best on τ-bench Airline drops 16 ranks once contributions are separated.

Report reliability with uncertainty. A score should come with a reliability estimate against the declared object and a credible interval. The minimum reportable set consists of the point estimate, credible interval, object of measurement, and estimation method. Without these, the leaderboard consequences in Section 3.3 are invisible to consumers.

Diversify contexts, not just task counts. Section 3.2 shows that single-benchmark model-ranking reliability saturates near $E\hat{\rho}^2 \approx 0.52$ under item-only scaling, because σ^2_{BM}, σ^2_{BMA}, and σ^2_{MA} do not attenuate with item count. Broadening across benchmarks attenuates the benchmark-conditioned terms and raises the projected ceiling to ≈ 0.88 at $n_b = 7$ in our considered benchmark set. The corresponding design prescription has three parts. First, allocate evaluation budget across independently informative benchmarks rather than concentrating it within one. Second, sample multiple

scaffolds per benchmark; current per-benchmark counts ($n_a \in \{2, 3\}$) leave σ^2_{MA} poorly estimated and absorbed into non-attenuating error. Third, report a D-study curve over n_i, n_a, and n_b so consumers can pick a design that meets a target reliability at minimum cost. The diversification budget should be set in proportion to the inference the leaderboard is intended to support: rankings of deployable model–scaffold systems require fewer scaffold draws than rankings of base-model capability across the task space.

D 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 [7, 23]. Our bounds are also within-benchmark: whether any object of measurement supports generalization *across* benchmarks, e.g., whether a model that ranks first on SWE-bench Verified Mini also ranks first on τ-bench Airline, is a separate validity question our framework does not directly answer. 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 item 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 Decomposition 4 partially addresses but per-benchmark scaffold-variance claims remain weaker, and pooling itself depends on HAL-Generalist as the cross-benchmark anchor (Appendix I.11). Future work should construct evaluation datasets that systematically span larger scaffold libraries, and should track how the ceiling shifts as the model pool evolves.

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 item selection by item discrimination, scaffold sampling under a defined scaffold family, and item-quality auditing; empirical evaluation of these, alongside methods that reduce the measurement cost of the recommendations in Section C, are the most direct extensions.

E Identifiability and estimation

The HAL-generalist scaffold is evaluated on every benchmark, providing a cross-benchmark anchor that identifies global model contrasts and prevents σ^2_M from being absorbed into benchmark- or scaffold-only terms; a formal connectivity

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

Benchmark	models	Tasks	Agents	model–Scaffold Pairs
AssistantBench	18	33	2	20
CoreBench-Hard	32	45	2	34
GAIA	21	165	2	23
SciCode	17	65	3	20
SWE-bench (Mini)	24	50	2	26
TAU-bench (Airline)	23	50	3	26
USACO	13	307	2	15

argument is in Appendix I.11. Main-text results use Bayesian logistic mixed models with weakly informative half-Student-t priors on standard deviations, fit via HMC in `brms`/Stan [**?**]; all parameters satisfy $\hat{R} < 1.05$. We replicate the analysis with LME, frequentist GLMM, and a nonparametric distance-components estimator (DISCO [22]); qualitative conclusions are robust across methods (Appendix K). Posterior model rankings use Best Linear Unbiased Predictors (BLUPs) of the model random effects.

F Data

We analyze agent rollouts from seven benchmarks distributed via HAL [14]: AssistantBench [31], CoreBench Hard [26], GAIA [17], SciCode [28], SWE-bench Verified Mini [12], τ-bench Airline [30], and USACO [25], covering web navigation, scientific reproducibility, multi-step assistance, scientific programming, software engineering, customer-service interaction, and competitive programming respectively (full descriptions in Appendix G).

Our analysis covers 20, 144 total agent rollouts. Each rollout is over a single task, model, reasoning-effort, and scaffold. Because our analysis is mostly interesting the relationship between the model-variant, scaffold and benchmarks, we join the reasoning-effort with the model to be our model-facet. For our analysis, each time model is stated it refers to the model at a specific-reasoning effort (i.e. 'gpt5-high' and 'gpt-5minimal' are seperated). Task counts range from 33 (AssistantBench) to 307 (USACO), model-Variant counts from 10 to 23, and each benchmark has 2 or 3 distinct scaffolds. All outcomes are task-level binary scores ($\text{score}_{ijk} \in \{0, 1\}$). The following benchmarks use a subset of tasks from the full benchmark, based on the available on the HAL leaderboard [13]: AssistantBench, GAIA, Scicode and SWE-bench Verified. Table 2 summarizes the design.

F.1 Data Limitations and Identifiability

Notice that every model–scaffold–benchmark combination is observed; the full coverage matrix is in Figure **??**. Several design features merit emphasis. First, the number of agent scaffolds per benchmark is extremely small (2–3), making scaffold variance difficult to estimate yet potentially consequential. Second, one scaffold—HAL-Generalist—appears in every benchmark, providing the critical thread of connectivity across the ecosystem and making it the linchpin of the pooled analysis used in Section 3.2-3.3 by anchoring model comparisons across otherwise disjoint evaluation pipelines and preventing global model variance from being an artifact of pooling. Care is taken to define Decomposition 4 that can be identified given the data. The design needs enough connectivity to separate "model signal" from "non-vanishing ecosystem heterogeneity," which HAL-Generalist provides. The pooled model borrows strength only because real overlaps exist (shared harness and shared models/items within benchmarks), so the decomposition is data-supported rather than purely model-imposed.

Furthermore, the findings are estimator- and method-robust. The same qualitative conclusions—nontrivial variance outside the model main effect implying limited model-ranking reliability—appears under Bayesian GLMM, frequentist GLMM/LME baselines, and nonparametric distance components, reducing concern about parametric or hierarchical artifacts. We report details for identifiability in Section I.11 and additional estimation method information in §I and K.

G Benchmark Descriptions

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.*

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...*

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 ($2 \le N \le 10^5$), 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.*

H Coverage Matrix

Below is the list of 2-3 agents used to test each of the benchmarks.

Benchmark	Agent Names
assistantbench	hal_generalist, assistantbench_browser_agent
corebench-hard	hal_generalist, coreagent
taubench-airline	hal_generalist, taubench_fewshot, taubench_tool_calling
swebench-verified-mini	hal_generalist, sweagent
scicode	hal_generalist, tool_calling_agent, scicode_zero
gaia	hal_generalist, hf_open_deep_research
usaco	hal_generalist, usaco_episodic_semantic

Table 3. Trace summary by benchmark.

Each model was not uniformly tested with each benchmark and agent. As show in Figure 5, about half the agents have only been tested on a single benchmark. Similarly, many model (11/30) are tested with a single agent.

The full breakdown of which models, agents and benchmarks have been tested can be seen in Figure**??**.

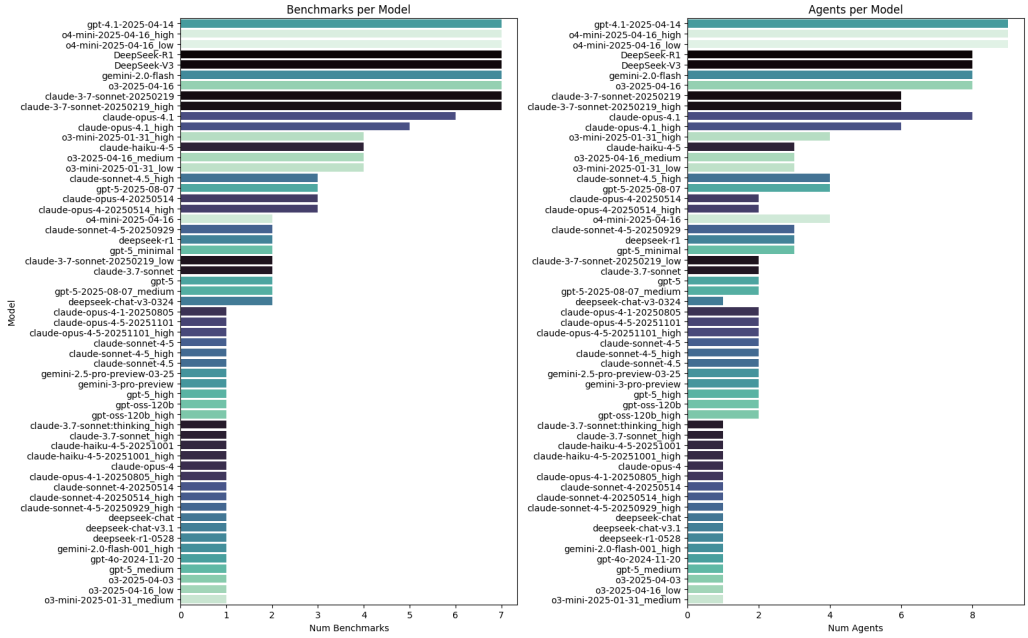

Fig. 5. Number of benchmarks tested (left) and agents tested (right) per model.

I Methods Details

I.1 Generalizability theory for agent benchmarks

G-theory can be used to explain the reliability of an experiment design by decomposing observed-score variability into variance components attributable to facets and their interactions. Reliability is defined as the fraction of observed-score variability attributable to the object of measurement rather than to sampling of other facets. In this section, we specify how we use G-theory to measure the reliability of claims made from agent evaluation data. First, we specify our facets, objects-of-measurement and formalizes a reliability metric (Section I.1.1), then propose four variance decompositions (Section I.1.2), and then specify our estimation procedure (Section I.1.3).

I.1.1 Facets, objects of measurement, and universe scores. Applying G-theory to an agent benchmark requires three commitments: the facets, the population each facet is sampled from (its *universe*, in G-theory terminology), and the target of inference (the object of measurement). The same data can yield different reliability estimates under different choices, so we make these commitments explicitly.

Facets. A score is indexed by benchmark b, item i, model m, and agent scaffold a. Let $y_{b,i,m,a} \in \{0, 1\}$ be the observed binary score. We treat all four facets as random, with sampling populations defined below.

Sampling populations. We define a narrow universe tied to the design realized in HAL, and discuss broader extrapolations in Section D.

- *Items* are treated as a random sample from the published item set of each benchmark. Generalization is to another draw from this set, not to the broader task domain the benchmark purports to represent (e.g., generalization is to

"items like SWE-bench Verified Mini," not to real-world software engineering). The construct-level question is a validity question, not a reliability question, as discussed in Section D.

- *Scaffolds* are treated as a random sample from the set of scaffolds in current evaluation practice for each benchmark. Per-benchmark scaffold counts are 2–3, reflecting a field-wide pattern in which most agent benchmarks are run under one or two scaffolds and the universe of evaluated scaffolds is itself a selected sample (see Section D). We pool across benchmarks in Decomposition 4 (Section I.1.2) to recover statistical power.
- *Models* are treated as a random sample from frontier models evaluated on HAL (Section F during the study period. This universe is clustered: many models share base architectures and training pipelines. Clustering depresses σ^2_M and is a feature of the current model landscape rather than an artifact of our measurement.

Object of measurement. Reliability is computed with respect to a chosen *object of measurement* o, the unit being ranked. We compute reliability for two objects, corresponding to two natural questions a leaderboard is used to answer: (i) ranking *models* ($o = m$), treating scaffold as a nuisance facet to be averaged over, and (ii) ranking *model–scaffold pairs* ($o = (m, a)$), treating the pair as the deployable unit and target of interest.

Relative reliability. Leaderboard use is a relative-decision problem: the question is which object ranks above which, not what each one's absolute score is. We therefore report relative reliability throughout, denoted $E\rho^2$, in which nuisance-facet main effects shared across all objects drop out and only object-by-facet interactions and residual terms enter the error. For an object o aggregated over n_i items and n_a scaffolds,

$$E\rho^2 \;=\; \frac{\sigma^2_o}{\sigma^2_o + \sigma^2_{\text{rel-error}}(n_i, n_a, \dots)}, \tag{4}$$

where $\sigma^2_{\text{rel-error}}$ collects object-by-facet interaction variances and the residual, each scaled by the relevant sample sizes; main effects of i, a, and b alone do not appear in the denominator. The function $\sigma^2_{\text{rel-error}}(\cdot)$ is the D-study (decision-study) projection: it describes how error shrinks as the number of sampled items, scaffolds, or benchmarks grows. The full expression for each decomposition is given in Appendix I.8. The framework extends straightforwardly to absolute reliability (Φ) for threshold decisions, e.g., certifying that a model exceeds a fixed performance bar [4].

I.1.2 Variance decomposition models. We estimate four variance decomposition models of increasing generalizability: three per-benchmark (Decomposition 1–3) and one pooled ecosystem model (Decomposition 4). Code and details for each are found in Appendix I.

Notation. Let $b \in \mathcal{B}$ index benchmark, $i \in \mathcal{I}_b$ index item (task), $m \in \mathcal{M}_b$ index model, and $a \in \mathcal{A}_b$ index agent scaffold. Let $y_{b,i,m,a}$ be the binary score when observed. The canonical GLMM is $y_{b,i,m,a} \sim \text{Bernoulli}(p_{b,i,m,a})$, $\text{logit}(p_{b,i,m,a}) = \eta_{b,i,m,a}$, with η expressed as a sum of random effects corresponding to facets and interactions (below). For LME fits, we treat y as approximately continuous and use an identity link; for GLMM and Bayesian GLMM, we use the logit link.

Decomposition 1 (naïve per-benchmark; object is model). This decomposition mirrors a common leaderboard assumption: scores vary by model and by item, with all other variability absorbed into residual noise. For each benchmark b: $\text{logit}(p_{i,m}) = \beta_0 + u^{(I)}_i + u^{(M)}_m$, with $u^{(I)}_i \sim \mathcal{N}(0, \sigma^2_I)$ and $u^{(M)}_m \sim \mathcal{N}(0, \sigma^2_M)$.

Object of measurement: model m. *Purpose:* quantify the optimism induced by omitting agent scaffold variance and model–agent interactions.

Decomposition 2 (per-benchmark; object is model–agent pair). This model redefines the object as the *model–agent pair*, absorbing agent and model×agent variability into "signal": $\text{logit}(p_{i,(m,a)}) = \beta_0 + u_i^{(I)} + u_{m:a}^{(MA)}$, where $u_{m:a}^{(MA)} \sim \mathcal{N}(0, \sigma^2_{MA})$.

Object of measurement: pair (m, a). *Interpretation:* reliability for ranking *deployable systems*, not bare models.

Decomposition 3 (per-benchmark; separates model, agent, and interactions). Model 3 is the per-benchmark design intended to reflect agentic evaluation reality: scores depend on model, item, agent harness, and their interactions:

$$\eta_{i,m,a} = \beta_0 + u_i^{(I)} + u_m^{(M)} + u_a^{(A)} + u_{i:m}^{(IM)} + u_{i:a}^{(IA)} + u_{m:a}^{(MA)}. \tag{5}$$

We treat the 3-way interaction $u_{i:m:a}^{(IMA)}$ as *confounded with residual* because several benchmarks lack repeated runs per (i, m, a), making separate identification unstable.

Objects of measurement: model m (primary) and, secondarily, pair (m, a). *Use:* compute D-study reliability under hypothetical changes in the number of items and/or agent scaffoldes, explicitly accounting for model–agent coupling.

Decomposition 4 (pooled ecosystem model across benchmarks). To use shared signal across harnesses and benchmarks (notably via HAL-generalist), we pool all benchmarks and include benchmark-level facets and cross-benchmark interactions:

$$\eta_{b,i,m,a} = \beta_0 + u_b^{(B)} + u_{b:i}^{(BI)} + u_m^{(M)} + u_a^{(A)} + u_{b:i:m}^{(BIM)} + u_{b:i:a}^{(BIA)} + u_{m:a}^{(MA)} + u_{b:i:m:a}^{(BIMA)} + u_{b:m}^{(BM)} + u_{b:a}^{(BA)} + u_{b:m:a}^{(BMA)}. \tag{6}$$

All random effects are modeled as mean-zero Gaussian with component variances to be estimated. This model supports the central question: *even if we scale item counts, is model-ranking reliability fundamentally bounded by cross-benchmark and harness variability?*

I.1.3 Estimation approaches. We fit Bayesian logistic mixed models with weakly informative priors on fixed effects and half-Student-t (or equivalent) priors on standard deviations: $\sigma_k \sim \text{Half-}t(\nu, 0, s)$, with conservative (ν, s) to regularize variance components away from degenerate solutions without dominating the likelihood. Posterior draws are used to compute posterior distributions over variance shares and reliabilities. All parameters across all models satisfy $\hat{R} < 1.05$. For estimated capability score for posterior ranking, Best Linear Unbiased Predictors (BLUPs) were extracted as draws of the estimated random effects. Each decomposition (1-4) is fit using three additional estimation methods to stress-test conclusions under small-n and non-Gaussianity. For Decomposition 4, we detail its identifiability in Appendix I.11 and demonstrate the stability global findings via four estimation method comparisons, all of which are detailed in Section I and Section K.

I.2 Model Specifications in Detail

We provide the full random-effects structure for each model using standard mixed-model notation. All models assume Bernoulli outcomes with logit link for Methods 2–4, and Gaussian identity link for Decomposition 1. We denote indices as: i for item (task), j for model (model name × reasoning-effort), k for agent scaffold, and b for benchmark.

I.2.1 Decomposition 1: Naïve.

$$\text{logit}\ \Pr(Y_{ij} = 1) = \mu + t_i + m_j, \quad t_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_t^2),\ m_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_m^2). \tag{7}$$

Fit per benchmark. The residual on the logit scale is the standard logistic variance $\pi^2/3 \approx 3.29$.

`brms` specification:

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

I.2.2 Decomposition 2: Model–Scaffold Pair.

$$\text{logit}\ \Pr(Y_{i,jk} = 1) = \mu + t_i + p_{jk}, \quad p_{jk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_p^2). \tag{8}$$

Fit per benchmark. The model×agent interaction is absorbed into the pair term p_{jk}.

brms specification:

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

I.2.3 Decomposition 3: Full Per-Benchmark.

$$\begin{aligned} \text{logit}\ \Pr(Y_{ijk} = 1) &= \mu + t_i + m_j + a_k + (tm)_{ij} + (ta)_{ik} + (ma)_{jk}, \\ t_i \sim \mathcal{N}(0, \sigma_t^2), &\quad m_j \sim \mathcal{N}(0, \sigma_m^2), \quad a_k \sim \mathcal{N}(0, \sigma_a^2), \\ (tm)_{ij} \sim \mathcal{N}(0, \sigma_{tm}^2), &\quad (ta)_{ik} \sim \mathcal{N}(0, \sigma_{ta}^2), \quad (ma)_{jk} \sim \mathcal{N}(0, \sigma_{ma}^2). \end{aligned} \tag{9}$$

The three-way interaction $(tma)_{ijk}$ is confounded with the residual in the absence of within-cell replication.

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"), ...)
```

I.2.4 Decomposition 4: Pooled Leaderboard.

$$\begin{aligned} \text{logit}\ \Pr(Y_{bijkl} = 1) = \mu + b_b + (bt)_{bi} + m_j + a_k \\ + (btm)_{bij} + (bta)_{bik} + (ma)_{jk} + (btma)_{bijk} \\ + (bm)_{bj} + (ba)_{bk} + (bma)_{bjk}. \end{aligned} \tag{10}$$

All random effects are independent and normally distributed. Items are nested within benchmarks, so t_i does not appear as a main effect; its role is absorbed by $(bt)_{bi}$.

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

I.2.5 Decomposition 4 Multivariate Extension. The multivariate extension models two outcomes simultaneously:

$$\begin{pmatrix} \text{logit Pr}(\text{score}_{bijkl} = 1) \\ \log(\text{tokens}_{bijkl}) \end{pmatrix} = \boldsymbol{\mu} + [\text{same random-effects structure as Model 4}], \tag{11}$$

with outcome-specific variance components and a residual correlation parameter ρ_ε between score and log-tokens.

`brms` specification:

```
brm(bf(mvbind(score, log(total_tokens)) ~ 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:task_id:model_name:agent_name)
    + (1|benchmark:model_name)
    + (1|benchmark:agent_name)
    + (1|benchmark:model_name:agent_name)), ...)
```

I.3 Bayesian Estimation Details

Priors. All models use `brms` default weakly informative priors: half-Student-$t(3, 0, 2.5)$ on random-effect standard deviations and a Student-$t(3, 0, 2.5)$ prior on the intercept. These priors regularize estimates away from boundary values ($\sigma = 0$) while remaining sufficiently diffuse to let the data dominate.

Sampling. Models 1–3 use 4 chains × 1,000 iterations with a thinning factor of 2, yielding 2,000 posterior draws per model. Model 4 uses 4 chains × 2,000 iterations with thinning of 5, yielding 1,600 draws. The target acceptance rate is set to `adapt_delta` = 0.95 across all models to reduce divergent transitions in the complex random-effects structure.

Convergence. All parameters across all models achieve $\hat{R} < 1.05$, the standard threshold for adequate mixing. We additionally inspect trace plots and effective sample sizes; the latter exceed 400 for all variance components.

Variance proportion computation. From each posterior draw $s = 1, \ldots, S$, we compute the variance share for facet f as

$$\pi_f^{(s)} = \frac{(\sigma_f^{(s)})^2}{\sum_g (\sigma_g^{(s)})^2 + \pi^2/3}, \tag{12}$$

where the sum runs over all random-effect standard deviations and $\pi^2/3$ is the logistic residual variance. Posterior summaries (mean, median, HDI) are computed over the $\{\pi_f^{(s)}\}$ draws.

I.4 Distance Components (DISCO)

The DISCO decomposition [22] generalizes classical ANOVA to arbitrary metric spaces. For K groups with n_k observations each, define the within-group dispersion as

$$\mathcal{S}_W = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i<j} \|X_{ki} - X_{kj}\|, \tag{13}$$

and the total dispersion as

$$\mathcal{S}_T = \frac{1}{2N} \sum_{i<j} \|X_i - X_j\|, \tag{14}$$

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.

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 item cluster—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.

I.5 Details On Compute-Usage

Estimation of all four methods across all four decompositions took 12 total hours on an Apple M1 Max.

I.6 Estimation Methods

Each of the four models is estimated by four methods, chosen to span the space from conventional to robust. In the main body, we report Method C. Findings are robust to estimation method.

I.6.1 Method A: 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.

I.6.2 Method B: 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 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` [2].

I.6.3 Method C: Bayesian Generalized Linear Mixed Effects. We fit the same logistic specification as Method B within a fully Bayesian framework using Hamiltonian Monte Carlo (HMC) via `brms`/Stan [5]. Weakly informative half-Student-t priors are placed on all standard-deviation parameters. Four chains of 1,000 (Models 1–3) or 2,000 (Model 4) iterations are run with thinning factors of 2 or 5, respectively, and `adapt_delta` = 0.95 to reduce divergent transitions. Convergence is assessed by $\hat{R} < 1.05$ for all parameters. Posterior means, medians, and 95% highest-density intervals (HDIs) of variance proportions are reported, providing a full uncertainty quantification absent from frequentist methods.

I.6.4 Method D: Distance Components (DISCO). We estimate a nonparametric variance decomposition using distance components [22, 27] 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,

$$\mathcal{S}_{\text{total}} = \mathcal{S}_{\text{between}} + \mathcal{S}_{\text{within}}, \tag{15}$$

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 variance shares sum to one across the total explained dispersion.

I.7 Reliability Estimation and Projection

From each fitted model, we extract the proportion of total variance (or dispersion, for DISCO) attributable to each facet. We additionally compute *projected* G-coefficients under hypothetical increases in the number of items, using the Spearman–Brown-like formula from G-theory:

$$E\rho^2(n'_I) = \frac{\sigma^2_m}{\sigma^2_m + \frac{1}{n'_I}\left(\sigma^2_{tm} + \sigma^2_\varepsilon\right)}, \tag{16}$$

where n'_I is the projected item count. For Model 4, analogous projections vary both item and benchmark counts.

I.8 Reliability Projection Formula

For a $p \times I$ design (models crossed with items), the projected G-coefficient under n_I items is

$$E\rho^2(n_I) = \frac{\sigma^2_m}{\sigma^2_m + \frac{\sigma^2_{mI} + \sigma^2_\varepsilon}{n_I}}, \tag{17}$$

where σ^2_{mI} is the model×item interaction variance. As $n_I \to \infty$, $E\rho^2 \to \sigma^2_m/\sigma^2_m$, but only if facets that do not diminish with more items (e.g., agent main effects, benchmark×model interactions) are negligible. When such facets are present, the asymptote is

$$\lim_{n_I \to \infty} E\rho^2 = \frac{\sigma^2_m}{\sigma^2_m + \sigma^2_{ma} + \sigma^2_{mb} + \cdots} < 1, \tag{18}$$

producing the reliability ceiling documented in our Model 4 analysis.

I.9 Variance Proportion Tables

Full variance-proportion tables for all four models and all four estimation methods are provided in the supplementary materials. Tables in the main text report Bayesian posterior summaries; the appendix tables additionally include LME, GLME, and DISCO estimates to enable cross-method comparison.

I.10 Latent Space Estimations

Variance components in our 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 for the mid-performing region typical of many AI-agent benchmarks (where $\mathrm{d}p/\mathrm{d}\eta = p(1-p)$ is largest), but its practical impact (if that were desired) should be assessed by mapping latent uncertainty through the inverse-logit.

Accordingly, in supplementary material, we report an additional observed-scale sensitivity analysis: for each posterior draw of variance components we (i) generate replicated item averages for each object under the fitted GLMM, (ii) transform to percent-correct $\bar{p}$ via logit^{-1}, and (iii) summarize rank variability (e.g., expected Kendall–τ or pairwise reversal probability) on the probability scale. This makes clear that our conclusions are not an artifact of a hard-to-interpret latent metric: the latent-scale ρ^2 is the 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.

I.11 Identifiability of the pooled ecosystem decomposition (Decomposition 4)

Decomposition 4 pools all benchmarks to estimate variance components for models, agent harnesses, benchmark effects, and their interactions. Because the evaluation matrix is sparse and unbalanced, we argue identifiability based on (i) *which parameters are identified from connectivity* in the observed design, (ii) *which are only weakly identified* (hence wide posteriors), and (iii) why the observed "structural ceiling" in model-ranking reliability is not a generic artifact of hierarchical shrinkage.

I.11.1 Design graph and why connectivity matters. Let an observation be indexed by a tuple $x = (b, i, m, a)$: benchmark $b \in \mathcal{B}$, item $i \in \mathcal{I}_b$, model $m \in \mathcal{M}$, and agent harness $a \in \mathcal{A}$. Write the linear predictor of Decomposition 4 as

$$\begin{aligned}\eta_{bima} = \beta_0 + u_b^{(B)} + u_{b:i}^{(BI)} + u_m^{(M)} + u_a^{(A)} + u_{b:i:m}^{(BIM)} + u_{b:i:a}^{(BIA)} \\ + u_{m:a}^{(MA)} + u_{b:i:m:a}^{(BIMA)} + u_{b:m}^{(BM)} + u_{b:a}^{(BA)} + u_{b:m:a}^{(BMA)},\end{aligned} \tag{19}$$

with $y_{bima} \sim \mathrm{Bernoulli}(\mathrm{logit}^{-1}(\eta_{bima}))$ and independent Gaussian random effects (mean zero, component variances σ_k^2). A convenient way to reason about identifiability in sparse crossed random-effects models is via a *connectivity graph*. Define a bipartite graph

$$G_b = (\mathcal{M} \cup \mathcal{A}_b,\, E_b), \qquad (m, a) \in E_b \;\Leftrightarrow\; \exists i \in \mathcal{I}_b :\; y_{bima} \text{ observed}.$$

Intuitively, if G_b has multiple disconnected components, then within benchmark b we cannot compare models across components because no agent harness provides a common measurement context. Decomposition 4 requires *cross-benchmark* connectivity as well. For that, define the global tripartite incidence structure induced by observed tuples (b, m, a):

$$(m, a, b) \text{ is present} \;\Leftrightarrow\; \exists i \in \mathcal{I}_b :\; y_{bima} \text{ observed}.$$

Role of HAL-generalist. One harness, denoted $a^\star$ (HAL-generalist), is evaluated on *every* benchmark and appears in all benchmarks that have only two harnesses. This induces a star-like connectivity pattern: for each b, many models satisfy $(m, a^\star) \in E_b$, and across benchmarks the same $a^\star$ provides a shared reference harness.

This matters because the decomposition includes both (i) *global harness effects* $u_a^{(A)}$ and (ii) *benchmark-conditioned harness effects* $u_{b:a}^{(BA)}$ as well as interactions with models. Without a harness present across benchmarks, $u_m^{(M)}$ and $u_{b:m}^{(BM)}$

(or $u_a^{(A)}$ and $u_{b:a}^{(BA)}$) can become only weakly separated: differences could be attributed either to model quality or to unobserved changes in harness composition.

I.11.2 Model identification.

(i) Identifiability of latent effects vs identifiability of variance components. Random effects are latent variables. In GLMMs, individual latent effects (e.g., a particular $u_{b:m}^{(BM)}$) are only identifiable up to their contribution to η, and can be weakly identified when few observations touch that level. What we require for reliability analysis is stronger: *identifiability of variance components* σ_k^2, since ρ^2 is a functional of $\{\sigma_k^2\}$ under a D-study.

(ii) Likelihood identifiability vs Bayesian propriety. A Bayesian posterior can be proper even when the likelihood is weakly identified, due to priors. Our goal is to argue that the key conclusion (a ceiling on model-ranking reliability) is supported by the *data connectivity*, not created by priors. Concretely: the data must support separating model variability from harness and benchmark variability, at least along the subspace needed to upper bound achievable ρ^2 as item count grows.

I.11.3 Anchoring and separating global vs benchmark-conditioned effects. We assume the standard random-effects centering constraints implied by the model specification:

$$\mathbb{E}[u^{(k)}] = 0 \quad \text{for all random-effect families } k,$$

which removes additive non-identifiability with the intercept. The remaining concern is *aliasing* among crossed terms, e.g., whether $u_m^{(M)}$ can be traded off with $u_{m:a}^{(MA)}$ or $u_{b:m}^{(BM)}$ without changing η on observed cells.

The HAL-generalist harness provides a measurement anchor that breaks key aliasing directions.

Lemma I.1 (HAL anchoring identifies global model contrasts across benchmarks). *Assume:*

1. *(**Cross-benchmark anchor**) There exists an agent harness $a^\star$ such that for every benchmark $b \in \mathcal{B}$, at least two distinct models $m_1, m_2 \in \mathcal{M}$ are observed with $a^\star$ on some items in benchmark b.*
2. *(**Within-benchmark item replication**) For each observed $(b, m, a^\star)$, scores are observed on at least two items (so that item effects do not perfectly saturate the cell).*

Then for any pair of models (m_1, m_2) that both appear with $a^\star$ in at least one common benchmark, the data identify their global *contrast $u_{m_1}^{(M)} - u_{m_2}^{(M)}$ up to benchmark-conditioned residuals that average out under item resampling; i.e., the contrast cannot be represented solely as a reallocation into agent-only or benchmark-only terms without changing η on observed cells.*

Sketch. Consider two observations in the same benchmark b and harness $a^\star$ but different models, and average across items:

$$\bar{\eta}_{b,m,a^\star} := \frac{1}{|\mathcal{I}_{b,m,a^\star}|} \sum_{i \in \mathcal{I}_{b,m,a^\star}} \eta_{bima^\star}.$$

Within the same $(b, a^\star)$, terms $u_b^{(B)}$, $u_{a^\star}^{(A)}$, and $u_{b:a^\star}^{(BA)}$ cancel in the difference $\bar{\eta}_{b,m_1,a^\star} - \bar{\eta}_{b,m_2,a^\star}$. Item-indexed terms shrink under averaging (and, in a D-study, under increasing item count). What remains includes $u_{m_1}^{(M)} - u_{m_2}^{(M)}$ plus model-conditioned interactions. Because $a^\star$ is fixed across both models, the model contrast cannot be fully absorbed into purely harness or benchmark main effects. □

Interpretation. Lemma I.1 is not claiming perfect analytic identifiability of every interaction term. Rather, it establishes that the observed design contains a repeated, shared harness that enables *global model comparisons* to be learned from data, which is the minimum needed to interpret σ^2_M and hence model-ranking reliability.

I.11.4 Observed ceiling would not be a hierarchical artifact. A common misunderstanding when interpreting larger hierarchical models is that when one such model is fit on sparse data, it will shrink everything and produce low reliability. In truth, shrinkage can increase or decrease estimated σ^2_M depending on how the data allocate variation among components. This is more clearly decided by whether the model is forced by its structure to attribute large mass to non-model components even when the data would support a high σ^2_M.

Any ceiling in Decomposition 4 would arise because it includes components that do *not* vanish with more items, and these components are empirically non-negligible under multiple estimators (LME, GLME, Bayesian, DISCO). *Even an infinite-item D-study cannot remove variability due to benchmark-conditioned model and harness interactions.*

Proposition I.2 (Non-vanishing error terms impose an upper bound on ρ^2_{model}). *Consider the pooled model with object of measurement m and a decision rule that aggregates scores by averaging over n_i items within each benchmark and a fixed harness mixture (as in observed leaderboards). Let σ^2_M be the global model variance and let σ^2_{nv} be the sum of variance components that are* not *attenuated by increasing n_i under this design (e.g., benchmark×model, model×agent, benchmark×model×agent, and other terms not scaled by $1/n_i$ in the D-study). Then the D-study model-ranking reliability satisfies*

$$\limsup_{n_i \to \infty} \rho^2_{\text{model}}(n_i) \;\le\; \frac{\sigma^2_M}{\sigma^2_M + \sigma^2_{nv}}.$$

In particular, if $\sigma^2_{nv} > 0$, reliability is structurally bounded away from 1 even with arbitrarily many items.

Sketch. In standard G-theory, averaging over n_i items attenuates components involving i by factors of $1/n_i$ (or $1/(n_i n_a)$ if also averaging over agents). Components that do not involve i (or that remain due to fixed harness mixtures and benchmark-conditioned shifts) remain. Reliability is signal over signal+error, hence the bound. □

Not a result of a larger measurement model. Proposition I.2 shows that the only way to avoid a ceiling is for the data to imply $\sigma^2_{\text{nv}} \approx 0$ (or to change the decision design, e.g., average over many harnesses and benchmarks). In our estimates, non-item components such as benchmark effects and benchmark-conditioned interactions remain appreciable across estimation methods, which is evidence that the ceiling reflects measured ecosystem heterogeneity rather than prior-driven shrinkage.

I.11.5 Connectivity is sufficient for the present conclusions. The paper's main conclusion is *not* that every variance component in Eq. (19) is precisely identified. Rather, it is that the ecosystem contains enough cross-context variability that model-ranking reliability cannot be made "high" by scaling item counts alone. For this, we need two claims:

Claim A: σ^2_M is not spuriously driven to zero. Because $a^\star$ appears across all benchmarks, global model contrasts are repeatedly observed under a common harness, preventing the model effect from being purely absorbed into benchmark or harness terms. This guards against the pathological case where σ^2_M is an artifactually small residual category.

Claim B: non-vanishing heterogeneity is data-supported. Benchmark-conditioned terms (e.g., $u^{(BM)}_{b:m}$, $u^{(BMA)}_{b:m:a}$) are supported by repeated observations within each benchmark across multiple models and at least one shared harness. In other words, the model is not inferring benchmark-conditioned interactions from a single cell; it is using within-benchmark comparisons and cross-benchmark anchoring via $a^\star$.

We summarize these two claims as:

Theorem I.3 (Sufficient connectivity for bounding model-ranking reliability). *Assume (i) there exists a cross-benchmark anchor harness $a^\star$ observed in every benchmark, (ii) within each benchmark at least two models are evaluated under $a^\star$ on multiple items, and (iii) at least one additional harness appears in some benchmarks (so that model×agent variability is not degenerate). Then:*

(1) *The global model variance σ_M^2 is estimable from repeated anchored contrasts and is not identifiable only through the prior.*
(2) *If the data support $\sigma_{nv}^2 > 0$ (in the sense that posterior mass for these components is away from zero under weakly informative priors and is corroborated by nonparametric DISCO decomposition), then Proposition I.2 implies a structural upper bound on ρ^2_{model} that persists as n_i grows.*

Discussion. Condition (i) is satisfied by HAL-generalist; condition (ii) is satisfied because each benchmark evaluates many models across many items under HAL-generalist; condition (iii) is satisfied because several benchmarks include additional harnesses (2–3 total), which is enough to reveal nontrivial $M \times A$ and benchmark-conditioned harness effects. Thus, while some high-order terms remain weakly identified (wide posteriors), the *existence* of non-vanishing heterogeneity–and hence a ceiling on ρ^2_{model} under item-scaling–is robust to modeling choices.

I.11.6 Practical takeaway for benchmark designers. To make model leaderboards stable, one must increase *connectivity*, not just sample size. The single most valuable design primitive is a *bridge harness* (like HAL-generalist) evaluated across benchmarks, plus enough additional harness diversity to estimate $M \times A$ variability. If a benchmark ecosystem lacks such anchors, pooled decompositions become prior-sensitive and any reliability claim becomes fragile. Here, the observed anchor connectivity ensures that the measured reliability ceiling reflects real cross-context variability in agentic evaluation rather than a hierarchical modeling artifact. This makes Decomposition 4 the strongest for interrogating the reliability of AI-agent benchmarking.

J Additional Plots

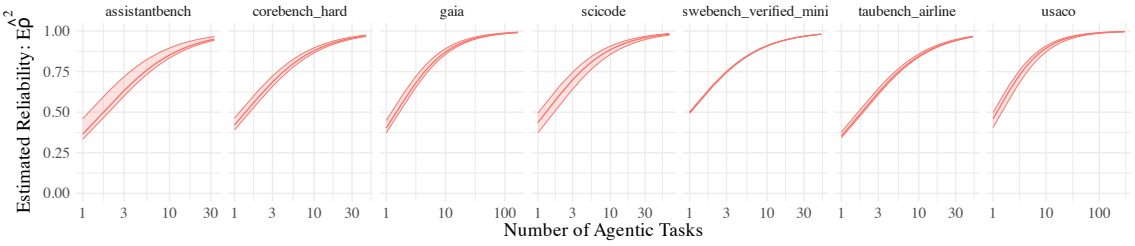

Fig. 6. Naïve Estimated Rank Reliability. Variance components aggregated after inverse logit transform, thus reliabilities represent the observation space.

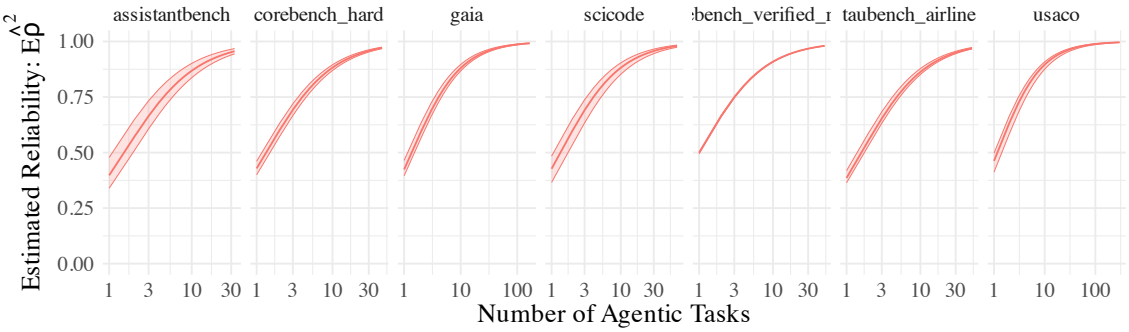

Fig. 7. Estimated Reliability of Rankings across Agentic Benchmarks, measured by Model-Scaffold pairing. Variance components aggregated after inverse logit transform, thus reliabilities represent the observation space.

Table 4. Rank and Distance Correlations between Original Rankings and Best Linear Unbiased Predictor Rankings

benchmark	Kendall	dcor	Spearman
assistantbench	0.699	0.762	0.863
corebench_hard	0.717	0.783	0.880
gaia	0.434	0.449	0.637
scicode	0.757	0.910	0.911
swebench_verified_mini	0.694	0.728	0.863
taubench_airline	0.380	0.154	0.488
usaco	0.779	0.646	0.915

For most benchmarks we see that general rank order (Spearman) and distribution are preserved (dCor) Rank and Distance Correlations between Original Rankings and Best Linear Unbiased Predictor Rankings

K Method Comparison

K.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.

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 items averaged (n_i items), an $I \times M$ component contributes σ_{IM}^2/n_i.

K.2 Distance components (DISCO) reliability

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

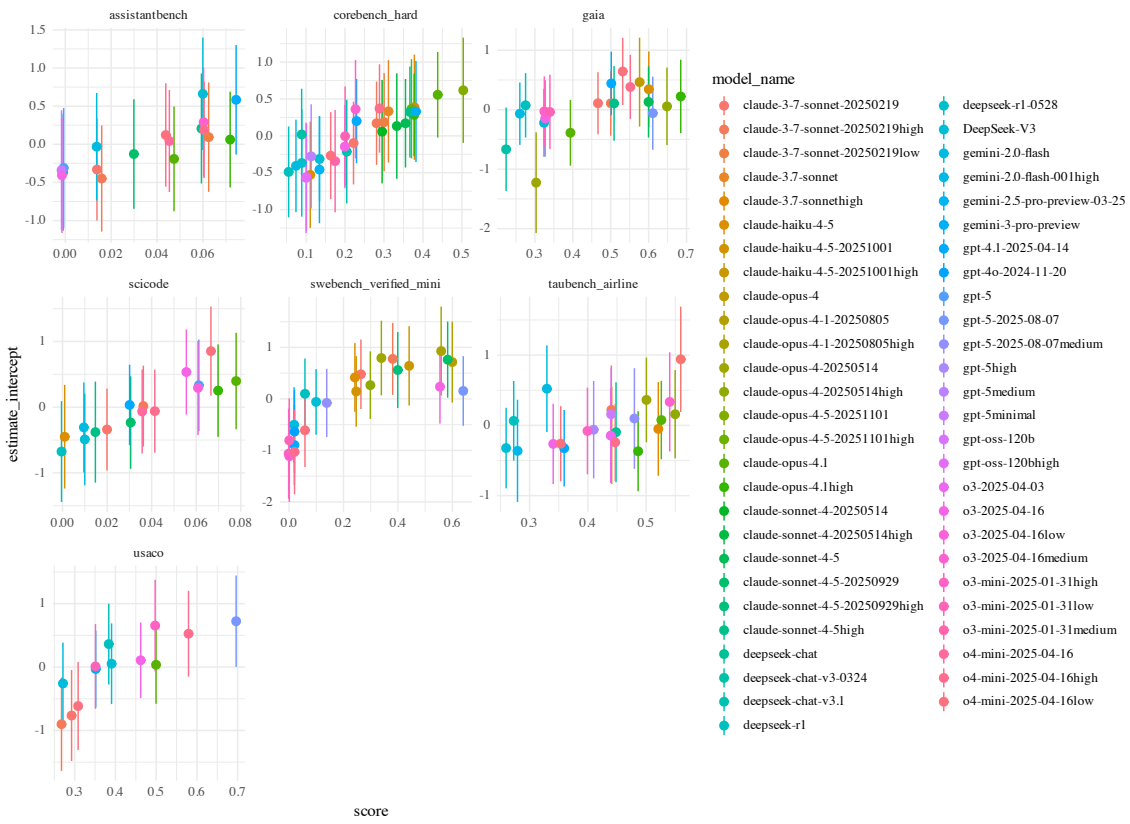

Fig. 8. Published scores and estimated scores with standard errors estimated from the Best Linear Unbiased Predictors from the posterior of Eq. 6

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 sensitivity analysis to validate that conclusions (dominant facets; low model-ranking reliability) are not artifacts of Gaussian random-effect assumptions.

K.3 Why estimation method choice changes conclusions—and what to do about it

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

A notable outcome is that variance shares differ across LME, GLME, Bayesian GLMM, and DISCO, especially under sparse observations:

- **LME** tends to allocate a large portion to residual σ^2, which can understate structured interactions when the link is misspecified for Bernoulli data.
- **GLME** can collapse some components toward zero (boundary estimates) in unbalanced designs, yielding deceptively "clean" decompositions.

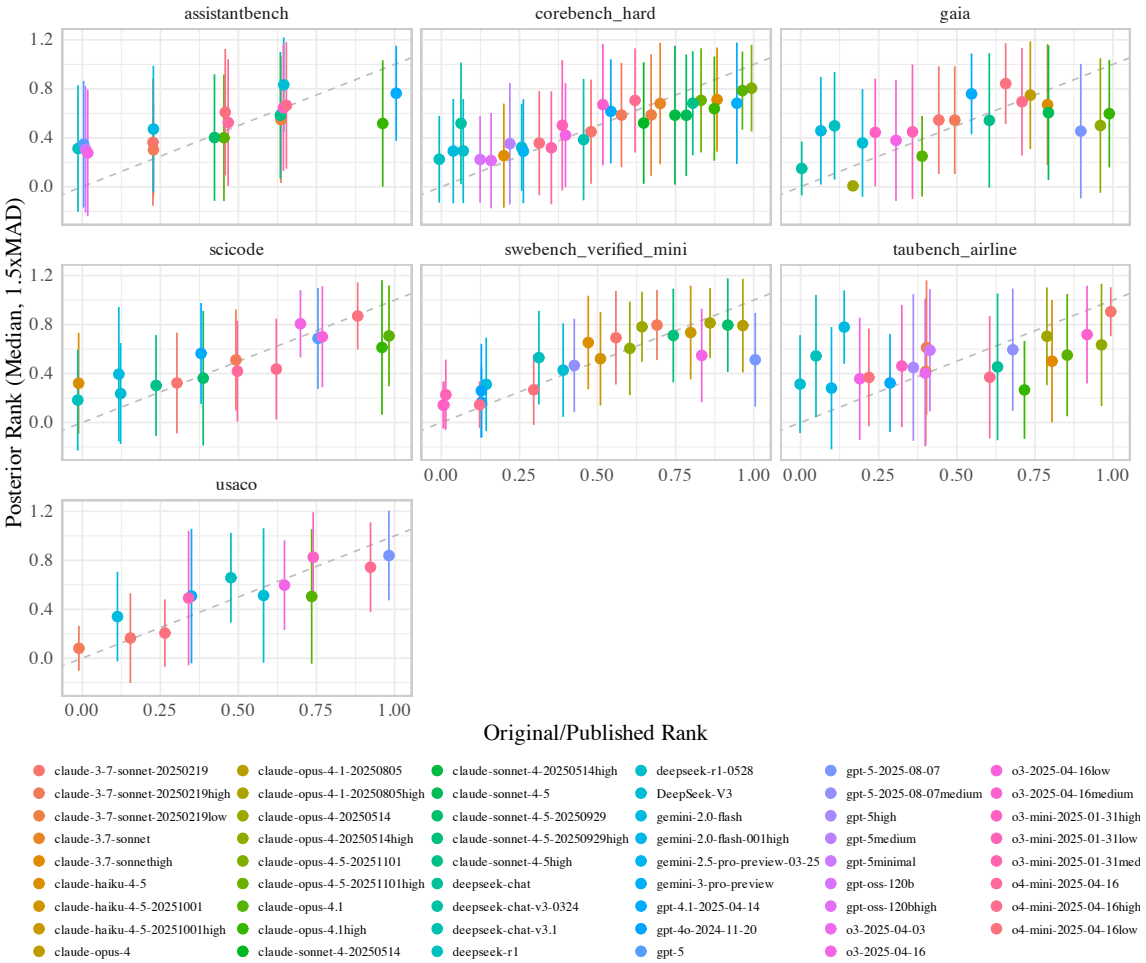

Fig. 9. Published rankings and draw-wise posterior median rankings of BLUPs with 90% CIs

- **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 GLME/LME only when coverage is dense.

(2) Triangulate: when all four methods agree on the *ranking* of dominant facets (e.g., item vs interactions vs residual), recommendations are robust; when they disagree, the correct conclusion is that the benchmark is under-instrumented for that inference.

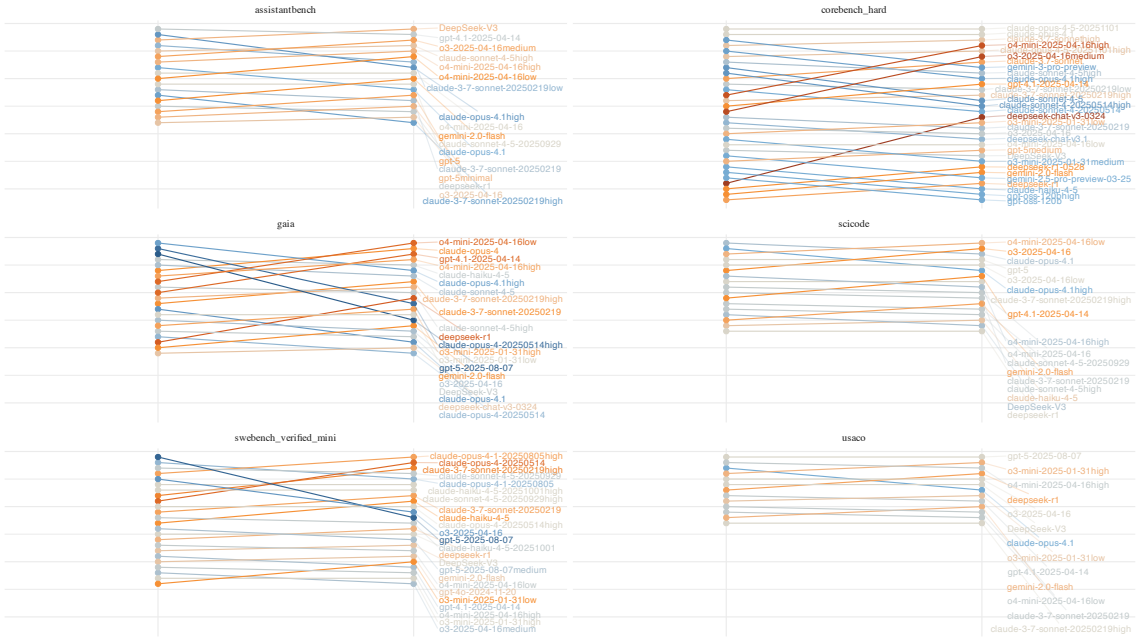

Fig. 10. Rank order changes by benchmark based on comparison with Best Linear Unbiased Predictors (BLUPs) from full variance decomposition model of Eq. 6. Taubench Airline can be found in the main body Figure 2

Linear vs. generalized models. The LME estimates consistently attribute the largest share of variance to the residual (45–85%), with correspondingly compressed named components. The GLME and Bayesian estimates, by contrast, allocate substantially more variance to items 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× 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. DISCO attributes 40–57% of total dispersion to item×model interactions across benchmarks, compared to 1–11% from the Bayesian estimates. This divergence likely reflects two mechanisms: (i) DISCO's sensitivity to nonlinear

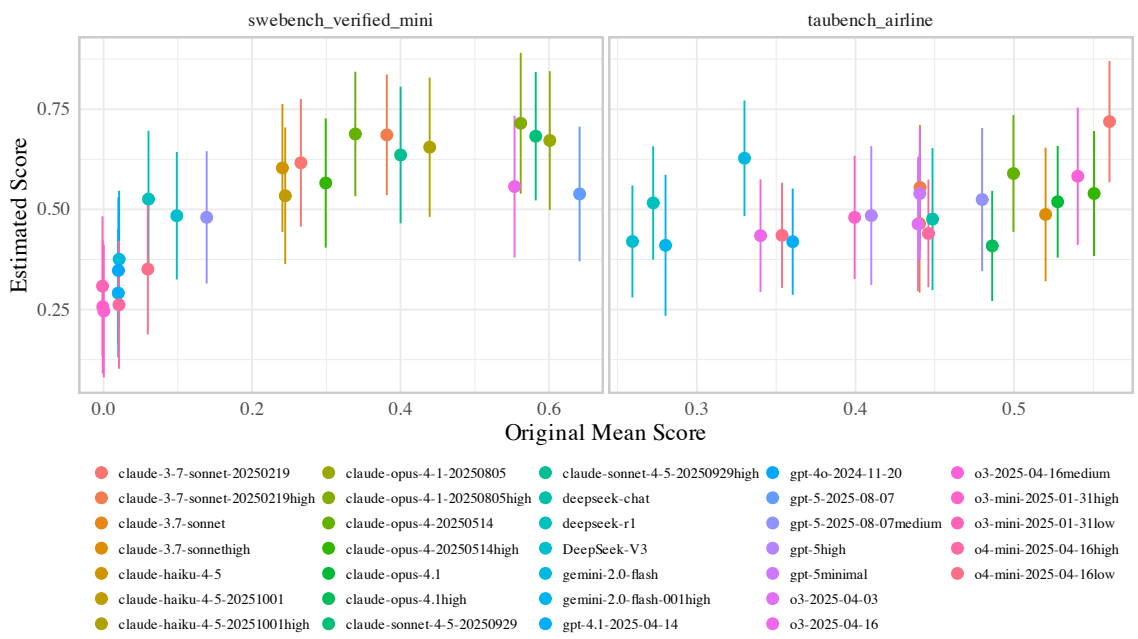

Fig. 11. Published scores and estimated scores with standard errors estimated from the Best Linear Unbiased Predictors from the posterior of Eq. 6. Estimates for all benchmarks are in Figure 8. Variance components aggregated after inverse logit transform, thus reliabilities represent the observation space.

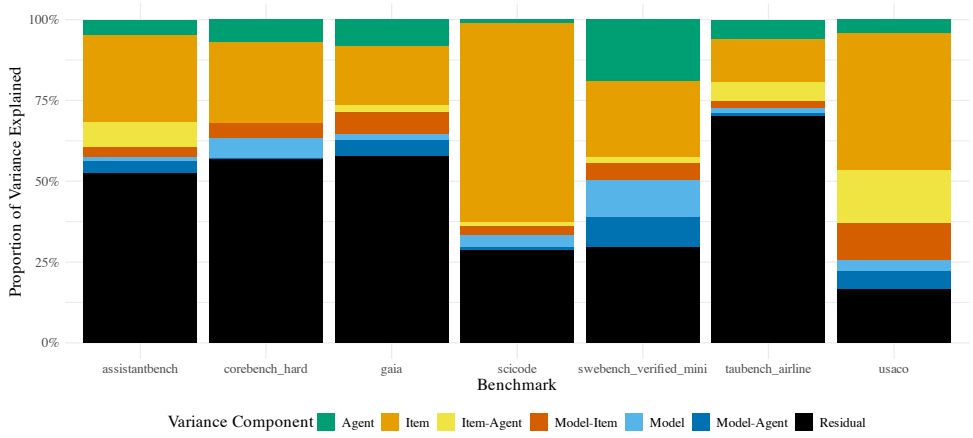

Fig. 12. Proportion of Variance Explained in the posterior distributions from Eq. 5.

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 model×item relationship.

K.4 Full Estimation Tables for Proportion of Latent Variation Explained

Table 5. Proportion of variation explained Decomposition 1

benchmark	facet	nonp	lme	glme	bayes
scicode	model	0.076	0.012	0.050	0.047
scicode	item	0.924	0.254	0.784	0.601
gaia	model	0.313	0.078	0.104	0.055
gaia	item	0.687	0.231	0.342	0.155
taubench_airline	model	0.268	0.025	0.033	0.016
taubench_airline	item	0.732	0.196	0.300	0.125
swebench_verified_mini	model	0.578	0.234	0.440	0.378
swebench_verified_mini	item	0.422	0.196	0.323	0.261
usaco	model	0.110	0.067	0.100	0.088
usaco	item	0.890	0.479	0.663	0.446
assistantbench	model	0.171	0.006	0.037	0.027
assistantbench	item	0.829	0.149	0.508	0.286
corebench_hard	model	0.293	0.073	0.121	0.067
corebench_hard	item	0.707	0.242	0.447	0.245
scicode	sigma		0.734	0.166	0.361
gaia	sigma		0.692	0.554	0.783
taubench_airline	sigma		0.780	0.667	0.841
swebench_verified_mini	sigma		0.570	0.237	0.597
usaco	sigma		0.454	0.237	0.492
assistantbench	sigma		0.845	0.455	0.678
corebench_hard	sigma		0.684	0.432	0.691

Received 20 May 2026

Table 6. Proportion of variation explained Decomposition 2

benchmark	facet	nonp	lme	glme	bayes
scicode	model	0.076	0.012	0.047	0.040
scicode	item	0.924	0.254	0.787	0.602
gaia	model	0.313	0.078	0.144	0.076
gaia	item	0.687	0.231	0.360	0.171
taubench_airline	model	0.268	0.025	0.092	0.041
taubench_airline	item	0.732	0.196	0.304	0.136
swebench_verified_mini	model	0.578	0.234	0.436	0.390
swebench_verified_mini	item	0.422	0.196	0.342	0.285
usaco	model	0.110	0.067	0.104	0.091
usaco	item	0.890	0.479	0.667	0.456
assistantbench	model	0.171	0.006	0.072	0.047
assistantbench	item	0.829	0.149	0.501	0.302
corebench_hard	model	0.293	0.073	0.134	0.072
corebench_hard	item	0.707	0.242	0.450	0.250
scicode	sigma		0.734	0.166	0.363
gaia	sigma		0.692	0.496	0.757
taubench_airline	sigma		0.780	0.605	0.809
swebench_verified_mini	sigma		0.570	0.222	0.574
usaco	sigma		0.454	0.230	0.482
assistantbench	sigma		0.845	0.427	0.649
corebench_hard	sigma		0.684	0.417	0.684

Table 7. Proportion of variation explained Decomposition 3

benchmark	facet	nonp	lme	glme	bayes
scicode	item	0.209	0.235	0.766	0.590
scicode	agent	0.001	0.001	0.004	0.055
scicode	model	0.012	0.010	0.053	0.046
scicode	item_agent	0.252	0.037	0.012	0.015
scicode	item_model	0.509	0.130	0.012	0.029
scicode	model_agent	0.017	0.002	0.003	0.011
gaia	item	0.169	0.210	0.342	0.182
gaia	agent	0.017	0.041	0.029	0.172
gaia	model	0.046	0.030	0.040	0.024
gaia	item_agent	0.209	0.036	0.035	0.024
gaia	item_model	0.481	0.097	0.041	0.068
gaia	model_agent	0.077	0.059	0.083	0.053
taubench_airline	item	0.161	0.166	0.277	0.134
taubench_airline	agent	0.021	0.043	0.044	0.123
taubench_airline	model	0.025	0.027	0.039	0.020
taubench_airline	item_agent	0.250	0.086	0.107	0.057
taubench_airline	item_model	0.484	0.035	0.000	0.021
taubench_airline	model_agent	0.059	0.012	0.016	0.011
swebench_verified_mini	item	0.092	0.153	0.349	0.261
swebench_verified_mini	agent	0.073	0.246	0.103	0.329
swebench_verified_mini	model	0.119	0.056	0.272	0.136
swebench_verified_mini	item_agent	0.190	0.068	0.024	0.023
swebench_verified_mini	item_model	0.399	0.089	0.009	0.057
swebench_verified_mini	model_agent	0.126	0.057	0.042	0.119
usaco	item	0.239	0.257	0.551	0.420
usaco	agent	0.003	0.072	0.000	0.102
usaco	model	0.027	0.062	0.000	0.048
usaco	item_agent	0.262	0.200	0.130	0.163
usaco	item_model	0.440	0.184	0.000	0.112
usaco	model_agent	0.030	0.000	0.103	0.060
assistantbench	item	0.156	0.116	0.437	0.254
assistantbench	agent	0.002	0.000	0.000	0.217
assistantbench	model	0.017	0.000	0.000	0.020
assistantbench	item_agent	0.219	0.083	0.107	0.090
assistantbench	item_model	0.574	0.047	0.000	0.044
assistantbench	model_agent	0.032	0.013	0.069	0.043
corebench_hard	item	0.172	0.239	0.446	0.251
corebench_hard	agent	0.007	0.016	0.019	0.175
corebench_hard	model	0.061	0.064	0.114	0.065
corebench_hard	item_agent	0.187	0.006	0.000	0.002
corebench_hard	item_model	0.502	0.111	0.036	0.046
corebench_hard	model_agent	0.071	0.009	0.007	0.006
scicode	sigma		0.584	0.150	0.283
gaia	sigma		0.527	0.431	0.571
taubench_airline	sigma		0.631	0.518	0.684
swebench_verified_mini	sigma		0.331	0.202	0.319
usaco	sigma		0.225	0.216	0.166
assistantbench	sigma		0.741	0.386	0.482
corebench_hard	sigma		0.555	0.378	0.556

Table 8. Proportion of variation explained Decomposition 4

facet	lme	glme	gbayes	nonp
agent	0.007	0.008	0.034	0.020
benchmark	0.097	0.287	0.268	0.030
benchmark_agent	0.035	0.032	0.026	0.037
benchmark_item	0.262	0.345	0.257	0.110
benchmark_item_agent	0.041	0.026	0.024	0.128
benchmark_item_model	0.063	0.011	0.033	0.229
benchmark_model	0.013	0.024	0.015	0.049
benchmark_model_agent	0.011	0.007	0.010	0.056
model	0.030	0.037	0.029	0.020
model_agent	0.008	0.013	0.009	0.043
benchmark_item_model_agent + sigma	0.433	0.210	0.304	0.278