How Stable Are Benchmark-Derived Difficulty Rankings?

Anonymous authors
Paper under double-blind review

Abstract

Analyses of benchmark data often identify which distinctions appear hardest, guiding model diagnosis, data curation, and subsequent evaluation. Yet a difficulty ranking is produced by a measurement procedure: fixed examples are represented, a supervised readout (probe) is fitted, and its scores are ordered. We construct pairwise discrimination tasks from benchmark labels and assess the stability of their ordering under representation and readout choices. On 400 AIOps2025 failure events spanning 55 mechanism pairs, we compare three representations and four fixed probes using grouped out-of-fold evaluation. Across all six probe contrasts under a fixed temporal representation, mean absolute rank shifts range from 3.69 to 14.55 positions; in the motivating logistic-regression–random-forest contrast, only three of the hardest ten pairs overlap. Numerical-feature, fit-seed and alternative grouped-fold controls place this displacement in context. RCAEval RE1 shows heterogeneous probe sensitivity across three distinct systems. Across six OpenML tabular datasets meeting prespecified eligibility criteria, median span-scaled rank displacement ranges from 0.065 to 0.233. Benchmark-derived pair difficulty is a measurable, procedure-relative output whose stability varies across settings and domains.

1 Introduction

Benchmarks do more than compare aggregate performance. Their labeled examples can be used to identify hard classes, hard distinctions, and subsets worth collecting or investigating next. For two failure types, pairwise separability offers one operational view: under a specified representation and probe, a lower out-of-fold classification score makes that distinction appear harder. Such analyses can influence what examples are selected for follow-up and what experimental capability is judged missing. These decisions use an ordering of tasks, not merely a mean score.

The ordering is itself a measured quantity. Even when benchmark examples and labels are fixed, a representation determines what aspects of an example are available to a learner; a probe converts that information into predictions; and a scoring rule orders the resulting pairwise tasks. A pair can consequently appear difficult under one reasonable measurement configuration and less difficult under another. This is distinct from ordinary random variation in fitting or splitting, and it is not resolved by asking which configuration obtains the best mean score.

We study the stability of benchmark-derived pair difficulty by holding a labeled cohort fixed and changing the representation or probe in controlled comparisons. AIOps/root-cause analysis (RCA) provides a deep testbed: failure types define natural pairwise distinctions, telemetry admits multiple summaries, and a second benchmark family supplies distinct systems. General numeric tabular benchmarks then test whether readout-dependent reordering occurs beyond RCA. The question concerns how an ordering derived from benchmark labels depends on the procedure producing it, rather than the accuracy of a particular RCA method.

The central issue is consequential reordering rather than the mere existence of score differences: its magnitude, which hard cases it selects, and whether it persists beyond fit-seed and fold variation.

On AIOps2025, 400 failure events yield 55 unordered mechanism pairs. Under one temporal representation and the same grouped folds, logistic regression and linear SVM agree closely (Spearman 0.947; mean absolute rank shift 3.69), while logistic regression and random forest rearrange the ordering

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.

(Spearman 0.459; shift 12.89; three of the hardest ten pairs overlap). Across three representations and four probes, the median pair spans 30 rank positions. A numerical feature check, repeated fit seeds, and two further valid grouped-fold assignments show which aspects of this pattern survive these changes. External comparisons span three RCAEval RE1 systems and six eligible OpenML tabular datasets, showing both recurrence and substantially different magnitudes across panels.

Our contributions are:

- A fixed-example measurement design that quantifies the stability of benchmark-derived pair rankings and hard-set membership across representation and supervised-readout choices.
- Controlled AIOps2025 evidence for structured representation and probe sensitivity, accompanied by numerical-feature, stochastic-fit, and valid grouped-fold controls.
- External recurrence across three RCAEval RE1 systems and six OpenML tabular datasets meeting prespecified eligibility criteria, revealing heterogeneous sensitivity within and beyond RCA.

2 Measuring Benchmark-Derived Difficulty

Let B_s be the fixed labeled example cohort for panel s, and let $\mathcal{P}_s$ contain all unordered pairs of its J_s class labels ($K_s = \binom{J_s}{2}$ pairs). The benchmarks supply examples and labels; our analysis constructs pairwise tasks and their ranking. A configuration $c = (r, h, m)$ specifies a representation r, probe h, and score m. For each pair $p = (a, b)$, the panel's predefined folds produce out-of-fold (OOF) predictions for precisely the examples with labels a or b. The pair score $S^c_{s,p}$ is computed from their pooled OOF outputs. Under the primary balanced-accuracy score, a lower value means less reliable separation by that configuration. We rank lower-scoring pairs as harder; equal scores receive average ranks, denoted $R^c_{s,p}$. The RCA folds preserve repeated fault/context groups; OpenML uses shared stratified folds.

For configurations c, c' applied to the same panel, we report Spearman and Kendall concordance and the mean absolute rank shift

$$\mathrm{MARS}_s(c, c') = \frac{1}{K_s} \sum_{p \in \mathcal{P}_s} \left| R^c_{s,p} - R^{c'}_{s,p} \right|, \qquad \mathrm{nMARS}_s(c, c') = \frac{\mathrm{MARS}_s(c, c')}{K_s - 1}. \tag{1}$$

Scaling MARS by $K_s - 1$, the full rank span of an individual pair, expresses average displacement relative to panel size. Appendix E checks a second scale based on the maximum whole-ranking footrule distance. We also compare deterministic hardest sets: pairs are ordered by $(S^c_{s,p}$, canonical pair ID) to resolve set-boundary ties. For within-AIOps hardest-set overlap, we set $q = 10$ from each full 55-pair ranking; cross-panel comparisons use $q_s = \lceil 0.30 K_s \rceil$, giving 17 AIOps2025 pairs, three RCAEval pairs per system, and dataset-specific sets of 5–17 pairs for OpenML. Overlap proportions and Jaccard, rather than raw intersection sizes, are compared across K_s.

Figure 1 distinguishes the fixed benchmark material from the varying measurement. The estimand is stability of this measured ordering among the specified configurations. It is not a latent ordering of intrinsic or Bayes task difficulty.

We organize the empirical study around three questions:

RQ1: Measurement sensitivity. How stable are pair-difficulty rankings when representation or probe changes while events and the primary folds are fixed?

RQ2: Stability controls. How do these changes compare with variation arising from a numerical-feature perturbation, stochastic fitting, and valid grouped-fold reassignment?

RQ3: External recurrence. Does measurement-dependent difficulty ordering recur across benchmark systems and outside the RCA domain?

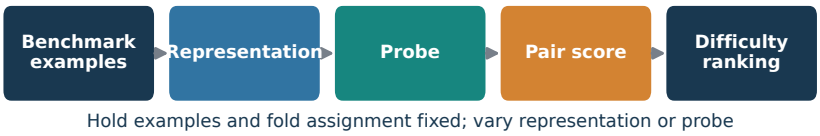

Figure 1: Fixed examples and fold assignments pass through alternative representations and probes to yield pair scores and an apparent difficulty ranking. Contrasts change one measurement component at a time.

3 Study Design

Cohorts and experimental unit. The unit is an unordered class/failure-type pair, scored on pooled OOF predictions from its labeled examples. AIOps2025 contributes the fixed 400-event cohort, 11 mechanisms and all 55 pairs. Three folds keep every mechanism-by-exact-context cell intact, so related events do not cross train and test. RCAEval RE1 contributes 375 metric-only case directories across Online Boutique, Sock Shop and Train Ticket (Pham et al., 2024). Two cases have injection timestamps outside the recorded telemetry interval and are excluded, leaving 373 eligible cases. All five faults and all five injected targets per fault remain represented in each system. The external panels are analyzed separately, never as one pooled ranking. Table 1 fixes the RCA denominators; Appendix D gives the OpenML panel definitions and exact dataset versions.

Table 1: RCA cohorts and evaluation designs. RCAEval counts follow source-time eligibility; its channel vocabulary is formed separately within each system. All folds group repeated fault–target/context cells.

Panel	Events	Types	Pairs	Channels	Representations
AIOps2025	400	11	55	104	Level, Change, Temporal
RE1 / Online Boutique	123	5	10	86	Level, Temporal
RE1 / Sock Shop	125	5	10	63	Level, Temporal
RE1 / Train Ticket	125	5	10	241	Level, Temporal

Representations. The three AIOps2025 summaries span complementary views of the same telemetry: interval levels, pre/post changes, and within-window temporal shape. Each of 104 base telemetry metrics has ten one-minute pre-event bins and ten equal-duration bins over the annotated failure interval. Each bin is the mean of finite observations. *Level* takes mean, median, population standard deviation, minimum and maximum over each ten-bin interval (ten features per metric). *Change* uses the differences in interval means, medians and standard deviations, the largest absolute failure-bin deviation from the pre mean, and a normalized mean difference with denominator $\max(|\bar{x}_{\text{pre}}|, 10^{-6})$ (five features per metric). *Temporal* applies the frozen 13-statistic per-metric transformation to the same 20 bins, retaining level, change and within-interval temporal descriptors. Nominal widths are 1,040, 520 and 1,352 respectively. No representation learns from labels during event-local feature construction.

The RE1 external comparison uses only Level and Temporal, with exact full Parquet column names as within-system channels and the same frozen bin/statistic functions. It uses ten bins over each complete observed pre-injection segment and ten over the recorded post-injection segment; unlike AIOps2025's fixed ten-minute pre period and explicit event end, RE1 has variable pre duration and no explicit fault-end timestamp. The post period is a faulty *observation* window, not an observed recovery interval.

Tabular cross-domain panels. Eight candidate multiclass numeric OpenML datasets (Vanschoren et al., 2013) were specified before fitting; six met the criteria of at least five classes and at least

30 examples in every class. Yeast and page-blocks were excluded because each had a class with fewer than 30 examples. The six eligible datasets (exact IDs/versions in Appendix D) produce 172 unordered class pairs. Each retains one original numeric representation and the same four fixed readouts. A fixed shuffled three-fold stratified assignment is shared across pair tasks and probes within a dataset; this is distinct from the RCA group-split rule. On all training-fold classes, numeric-column availability/variance filtering, mean imputation and scaling are fitted before pair selection and applied to held-out rows. Each pair's pooled OOF BA is primary and macro-F1 is a same-prediction check. All six probe contrasts enter each dataset median.

Probes, folds, and scores. Four fixed probe specifications are balanced logistic regression ($C = 1$, `lbfgs`), balanced dual linear SVM ($C = 1$), balanced RBF SVM ($C = 1$, `gamma=scale`), and a balanced-subsample random forest (100 trees, depth eight, minimum leaf two). These cover linear probabilistic and margin readouts alongside nonlinear kernel and tree-ensemble readouts. Fixed hyperparameters avoid adding per-system or per-pair tuning as another measurement component. Additional fixed parameters are in Appendix A. Within each representation/fold, feature availability (≥ 0.90), variance ($> 10^{-12}$), mean imputation and standardization are fitted on training events *before* pair selection and then applied to held-out events. This gives all binary pairs in a fold the same train-only feature map, avoiding pair-specific changes in feature availability and scaling. Each binary pair has both classes in every train and test partition. Balanced accuracy (BA) is primary, macro-F1 uses the same hard OOF predictions as a robustness score, and AUROC is a secondary AIOps sensitivity measure using probability/decision scores without metric-specific refitting. Pooled SVM margins are not probability-calibrated across folds.

Contrasts and uncertainty. RQ1 changes representation with probe and metric fixed, or probe with representation and metric fixed; all three AIOps representations, four probes, and 55 pairs are retained. RQ2 uses the same pair population: a feature-only Change-Core check, five additional fit seeds for the two stochastic frozen implementations, and two further prespecified valid grouped-fold assignments. RQ3 compares Level/Temporal and all four probes within each RE1 system using fault-by-target-cell grouped folds, and compares all six readout pairs within each OpenML dataset using its one numeric representation. Paired, class-stratified cell-cluster bootstrap intervals for RCA and within-class cached-prediction bootstrap intervals for OpenML use 10,000 common resamples per panel across compared configurations, without refitting or fold reassignment. These intervals are conditional on observed cohorts and trained OOF predictions.

The initial fixed-Temporal logistic-versus-forest comparison was exploratory. The expanded AIOps measurement and control protocols and the RCAEval/OpenML protocols were specified before their respective outcome analyses; Change-Core was a post-hoc numerical sensitivity check.

4 Measurement Sensitivity on AIOps2025

4.1 Probe choice under a fixed representation

Temporal Summary exposes substantial but non-uniform differences among the four fixed probes (Table 2). LR and linear SVM yield Spearman 0.947 and MARS 3.69 over 55 pairs. LR and RBF SVM shift 11.40 ranks on average; LR and RF shift 12.89 (Spearman 0.459), with only three pairs shared by their hardest-ten sets. RBF SVM and RF have MARS 14.55. Table 2 reports all six Temporal probe contrasts; the supplement gives the remaining representation/probe comparisons. The contrasts reveal relatively close and substantially different measurement configurations.

The LR–RF result also illustrates why a mean score difference cannot substitute for rank analysis. The pair-macro RF-minus-LR BA gain is 0.0467, while the mean absolute pairwise score difference is 0.0986 and includes changes in both directions. A user selecting the hardest ten mechanism distinctions would therefore obtain different follow-up cases even under the same representation, cohort and folds. The conditional cluster-bootstrap interval for their MARS is $[10.29, 16.02]$ positions; this interval conditions on the cached OOF procedure.

Table 2: AIOps2025 BA contrasts. All six Temporal probe comparisons appear individually; representation rows report medians across four fixed probes, including the median of four integer hardest-ten intersection counts. MARS averages absolute rank displacement across all 55 pairs. Per-probe representation contrasts and metric checks appear in the supplement.

Controlled comparison	Spearman	MARS (positions)	Hardest-10
Probe changes—individual contrasts (Temporal Summary fixed)			
LR–linear SVM	.947	3.69	6
LR–RBF SVM	.483	11.40	6
LR–RF	.459	12.89	3
linear SVM–RBF SVM	.516	11.67	5
linear SVM–RF	.448	13.60	4
RBF SVM–RF	.282	14.55	5
Representation changes—medians across four fixed probes			
Level–Change	.488	12.25	5
Level–Temporal	.792	7.58	7
Change–Temporal	.686	9.04	6.5

4.2 Representation choice and pair-level structure

With the probe held fixed, the median BA MARS over the four probes is 12.25 for Level–Change, 7.58 for Level–Temporal, and 9.04 for Change–Temporal (Table 2). Across the 12 representation–probe BA configurations, each pair has a range of measured ranks: the median pair range is 30 positions and the largest is 51. Figure 2 shows every pair's minimum, quartiles, median and maximum, preserving low- and high-range pairs instead of displaying only striking examples. The analogous macro-F1 contrasts broadly reproduce the pattern; at fixed representation and probe, BA versus macro-F1 rankings shift by a median 1.96 positions. AUROC produces more distinct ranks (median BA–AUROC shift 5.12) and remains secondary because pooled SVM decision margins are not calibrated as cross-fold probabilities.

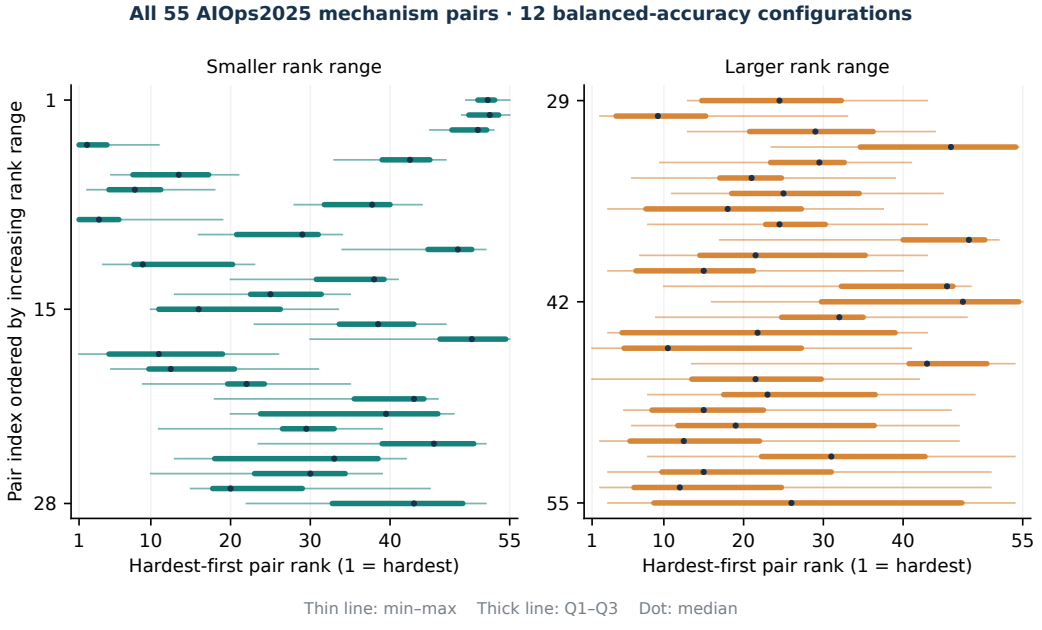

Figure 2: All 55 pair-rank distributions across 12 AIOps2025 BA measurement configurations, ordered by rank range. Thin segments show min–max, thick segments Q1–Q3, and dots the median. Pair-index mapping and exact values are in the supplement.

5 Stability Controls

The original Change representation contains one normalized mean-difference feature per metric whose finite magnitude reaches approximately 2.52×10^{12}. A post-hoc numerical check removes *only* this normalized component, preserving the four other Change features. Its BA median MARS remains 11.73 for Level–Core and 8.13 for Core–Temporal, compared with 12.25 and 9.04 for the corresponding original-Change contrasts. Original Change versus Core itself shifts 3.90 ranks on average across probes. The extreme feature contributes to the measurement, but the observed representation-associated rearrangement is not confined to it.

Under fixed Temporal Summary and the primary grouped folds, fit-seed comparisons yield BA MARS zero for the dual linear SVM across all 15 tested condition pairs, and median 2.15 (maximum 2.95) for RF. The latter retains a median of nine of the hardest ten pairs. These same-procedure changes are smaller than several of the between-probe contrasts observed in this cohort, though not all. Two further prespecified grouped-fold assignments produce same-probe, between-fold median BA MARS 4.44 (range 3.13–6.64); within each fold assignment, between-probe median MARS over the 18 comparisons is 12.45. The LR–RF shift is $12.89/13.15/11.85$ on primary/alternate-1/alternate-2 folds, and RBF SVM–RF is $14.55/14.95/15.13$. Valid reassignment changes individual ranks, but the larger probe contrasts persist under both alternatives.

Figure 3 presents all contrasts within these comparison families. Their observed ranges overlap; the plot compares controlled scales rather than partitioning variation into causal components.

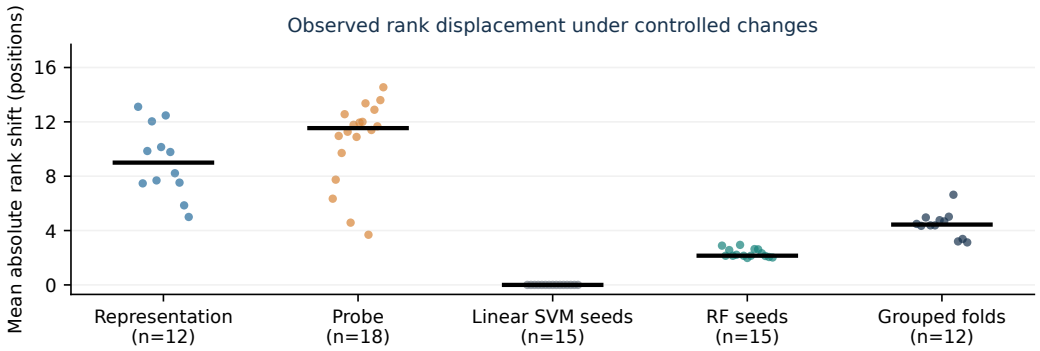

Figure 3: AIOps2025 BA displacement for every specified representation, probe, fit-seed and between-fold contrast shown here. Each shift averages over all 55 mechanism pairs; black bars mark reported group medians. Contrast groups reuse the same pairs but change different measurement components; they are descriptive comparisons rather than additive variance components.

6 External Recurrence

6.1 RCAEval RE1: recurrence across RCA systems

The RCA external study holds each system's eligible cases and grouped folds fixed while comparing Level with Temporal under each probe and all six probe pairs within each representation. Each system has five fault types and ten pair rankings. We report all four representation and twelve probe contrasts per system, with BA primary and macro-F1 as a same-output check. Table 3 reports median span-scaled nMARS, hardest-30% Jaccard and Spearman across all contrasts in each family. Figure 4 displays their individual contrasts beside the six OpenML panels.

Table 3: RCA BA ranking stability. Medians across four same-probe Level–Temporal contrasts (Rep.) or twelve within-representation probe contrasts (Probe). AIOps2025 provides contextual scale, not a matched-design comparator. Each row ranks only pairs within its own system; the full 64 individual rows are supplied separately.

Dataset/system	K	Contrast	Spearman	nMARS	Hard-30% Jaccard
AIOps2025	55	Rep.	.79	.14	.51
		Probe	.53	.21	.36
RE1 / Online Boutique	10	Rep.	.74	.16	.75
		Probe	.54	.22	.50
RE1 / Sock Shop	10	Rep.	.84	.12	.50
		Probe	.54	.23	.50
RE1 / Train Ticket	10	Rep.	.93	.11	.75
		Probe	.90	.09	.50

Online Boutique. Across its twelve probe comparisons, median nMARS is 0.217 (range 0.067–0.367). The Level LR–linear-SVM comparison is relatively close (Spearman 0.913, nMARS 0.067), whereas Level LR–RBF SVM differs markedly (Spearman 0.086, nMARS 0.333, one of three hardest pairs shared). The four representation contrasts have nMARS 0.100–0.178. Thus the external sensitivity has structure even within one system.

Sock Shop. Median probe nMARS is 0.233 (0.122–0.311); Level LR–RBF SVM reaches Spearman 0.328 and nMARS 0.311. Its Level–Temporal contrasts span nMARS 0.067–0.178. The broad movement is similar in scale to Online Boutique without requiring the same individual pair ordering.

Train Ticket. This panel is more stable under the specified procedures: median probe nMARS is 0.094 (0.067–0.189), with median probe Spearman 0.90. The Level–Temporal nMARS range is 0.022–0.167. Train Ticket shows that measurement sensitivity is not uniform across systems. Macro-F1 preserves the broad panel ordering: median probe nMARS is 0.22, 0.26 and 0.09 in Online Boutique, Sock Shop and Train Ticket respectively.

The paired fault-by-target-cell bootstrap uses 10,000 common resamples per system and cached predictions. For example, OB Level LR–RBF SVM has a conditional nMARS interval $[0.189, 0.444]$, while TT's RF Level–Temporal interval is $[0, 0.078]$. The complete supplementary interval tables include stable and unstable contrasts. Across all twelve probe contrasts, median nMARS is 0.210 for the AIOps2025 contextual comparison and $0.217/0.233/0.094$ for the RE1 panels. Scaling instead by the maximum whole-ranking footrule distance slightly reverses the close AIOps2025/Online Boutique comparison while preserving the marked contrast with Train Ticket (Appendix E).

6.2 OpenML: recurrence beyond RCA

Readout-dependent class-pair ordering is present outside the RCA domain, with substantially different magnitude among eligible datasets (Table 4). Across six fixed readout contrasts per dataset, median BA nMARS ranges from 0.065 (segment) to 0.233 (gas-drift). Pendigits (0.176) and texture (0.156) also show larger movement, whereas steel-plates-fault (0.076) and satimage (0.071) are comparatively stable. The macro-F1 column shows the same broad pattern using the same OOF hard predictions. Figure 4 shows all six individual contrasts together with each dataset median. Appendix D specifies dataset versions and resampling; complete BA/macro-F1 contrast and conditional bootstrap tables are in the supplement.

6.3 Cross-benchmark synthesis

The AIOps2025 probe median nMARS (0.210) and RE1 Online Boutique/Sock Shop/Train Ticket medians ($0.217/0.233/0.094$) offer contextual displacement scales beside the OpenML range (0.065–0.233). AIOps2025 and RE1 probe medians span multiple representations; OpenML has one numeric representation and different fold semantics. Thus the common observation is recurrence of measurement-dependent ordering with heterogeneous magnitude, not equality of effects under matched designs.

Table 4: All six eligible OpenML panels. Each entry is the median over all six fixed readout contrasts within one numeric representation; BA is primary and macro-F1 uses the same OOF predictions. K is the number of unordered class pairs; ρ denotes Spearman rank correlation; Jaccard compares hardest $\lceil .30K \rceil$ sets.

Dataset	K	BA ρ	BA nMARS	BA fMARS	Hard Jaccard	F1 nMARS
steel-plates-fault	21	.921	.076	.145	1.000	.090
gas-drift	15	.532	.233	.438	.429	.224
segment	21	.937	.065	.125	.875	.065
satimage	15	.934	.071	.134	.667	.067
pendigits	45	.640	.176	.344	.601	.174
texture	55	.676	.156	.307	.589	.157

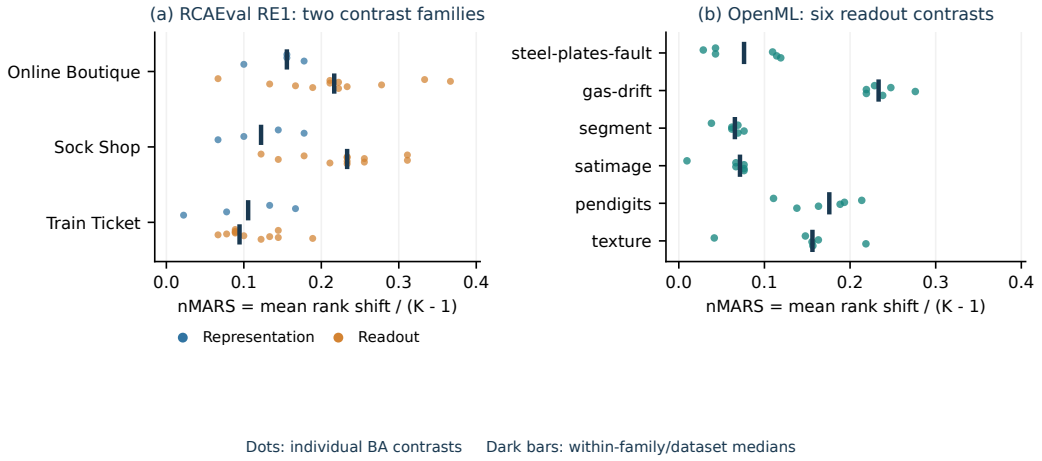

Figure 4: External BA recurrence in separate panels. Left: four representation and twelve readout contrasts per RCAEval system, with within-family medians. Right: all six readout contrasts and their medians for every eligible OpenML dataset. The horizontal nMARS scale normalizes by each panel's $K - 1$; fold and representation designs differ across benchmark families.

7 Discussion

The results locate a concrete sensitivity in a common use of benchmarks. A ranking can guide which distinctions receive more data, specialized modeling, or diagnostic attention. In AIOps2025, holding examples and folds fixed does not uniquely determine that ranking: some representations and probes agree closely, while others change the pair order and the hardest-set membership. A reported hard subset is therefore most informative together with the measurement configuration that produced it.

The controls test three narrower explanations of the observed movement. Removing the numerically extreme relative-change component leaves substantial representation contrasts. Ordinary stochastic fit seeds produce smaller rank shifts than several probe comparisons in this setting. Two new valid grouped splits alter ranks but preserve several larger probe contrasts.

Sensitivity is structured rather than universal. Online Boutique and Sock Shop exhibit substantial probe-dependent rearrangement while Train Ticket is more stable under the same four fixed probe specifications. The six OpenML datasets extend the recurrence outside RCA, with pronounced movement in gas-drift, pendigits and texture alongside comparatively stable steel-plates-fault, segment and satimage. When hard-case rankings determine data collection, benchmark curation or diagnostic follow-up, a small audit across plausible readouts can reveal whether a selected hard subset is procedure-sensitive. Contrasts and hard-set overlap add information that a mean score or a single probe alone cannot.

8 Related Work

Difficulty of examples and datasets. Example forgetting tracks how often an individual training item becomes misclassified during learning and identifies patterns that can persist across architectures (Toneva et al., 2019). Dataset cartography uses training confidence and variability to distinguish easy, hard and ambiguous examples relative to a model (Swayamdipta et al., 2020). A complementary account defines dataset and pointwise difficulty relative to a predictive model class through usable information (Ethayarajh et al., 2022). These studies establish that difficulty need not be model-free. Our unit is instead a *pairwise discrimination task* built from fixed benchmark labels, and our outcome is how the ordering among those tasks changes across measurement specifications.

Benchmark measurement and ranking sensitivity. Construct-validity work examines how an operational benchmark score supports its intended interpretation (Salaudeen et al., 2025; Bean et al., 2025). Evaluation perturbations can change model leaderboard order, including through answer-choice and scoring details (Alzahrani et al., 2024). That line studies the ranking of *models*; we hold the benchmark cohort fixed and study the ranking of *tasks* derived from it. The distinction matters when a hard-case list, rather than a winning model, determines the next data-collection or diagnostic priority.

Probes and benchmark domains. Linear probes and control tasks show why a supervised readout's capacity affects what can be inferred from its score (Alain & Bengio, 2016; Hewitt & Liang, 2019). We compare four fixed probe specifications as alternative measurements, without selecting a winning readout. RCA benchmark and diagnosis studies supply labeled failure mechanisms and a multi-system setting (Pham et al., 2024; Zhang et al., 2025a,b; Fang et al., 2026). The cross-domain tabular panels ask whether the resulting task-order sensitivity is peculiar to RCA telemetry or recurs with ordinary numeric features and class labels.

9 Limitations

Pairwise discrimination is one operationalization of difficulty; one-versus-rest, multiclass and generative tasks may order hard cases differently. The tested settings comprise AIOps2025, three RCAEval systems and six eligible numeric OpenML datasets from a prespecified non-probability candidate set, not image/text benchmarks or learned-representation replications beyond AIOps. Readouts and hyperparameters are fixed. Conditional bootstrap intervals reuse fitted OOF predictions rather than refitting models. RCAEval's variable pre/post windows and absent explicit fault-end time differ from AIOps2025's annotated intervals; OpenML uses stratified rather than group-preserving folds. A higher-iteration numerical sensitivity check for warned OpenML Linear-SVM fits leaves the cross-dataset pattern unchanged (Appendix D.1). The estimates describe measurement-relative rankings, not intrinsic or Bayes difficulty.

10 Conclusion

With benchmark examples fixed, changes in representation and probe can alter which pairwise distinctions appear difficult. AIOps2025 shows structured movement across 55 pairs that persists through numerical, fit-seed and grouped-fold controls. RCAEval RE1 shows heterogeneous recurrence across systems, while six eligible OpenML datasets show heterogeneous recurrence beyond RCA. Difficulty rankings used for benchmark interpretation and follow-up selection are empirical outputs of a stated measurement procedure whose stability can be checked.

AI use statement

Generative AI tools assisted literature organization, code development, and manuscript drafting and editing. The authors are responsible for the scientific analyses, interpretations, and final text.

Reproducibility statement. The appendix gives implementation details, dataset versions and scaling sensitivity; the anonymous supplement contains complete pair-level and contrast tables, conditional intervals, method summaries, a validation summary and numerical sensitivity results (Appendix F).

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. URL https://arxiv.org/abs/1610.01644.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 13787–13805, 2024. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2511.04703.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008, 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Aoyang Fang, Songhan Zhang, Yifan Yang, Haotong Wu, Junjielong Xu, Xuyang Wang, Rui Wang, Manyi Wang, Qisheng Lu, and Pinjia He. Rethinking the evaluation of microservice RCA with a fault propagation-aware benchmark, 2026. URL https://arxiv.org/abs/2510.04711.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Luan Pham, Huong Ha, and Hongyu Zhang. Root cause analysis for microservice systems based on causal inference: How far are we? In *IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2024. doi: 10.1145/3691620.3695065.

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, 2025. URL https://arxiv.org/abs/2505.10573.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9275–9293, 2020. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL https://doi.org/10.1145/2641190.2641198.

Shenglin Zhang, Xiaoyu Feng, Runzhou Wang, Minghua Ma, Wenwei Gu, Yongqian Sun, Zedong Jia, Jinrui Sun, and Dan Pei. Too many cooks: Assessing the need for multi-source data in microservice failure diagnosis. In *2025 IEEE 36th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 1–12, 2025a. doi: 10.1109/ISSRE66568.2025.00014.

Shenglin Zhang, Sibo Xia, Wenzhao Fan, Binpeng Shi, Xiao Xiong, Zhenyu Zhong, Minghua Ma, Yongqian Sun, and Dan Pei. Failure diagnosis in microservice systems: A comprehensive survey and analysis. *ACM Transactions on Software Engineering and Methodology*, 2025b. doi: 10.1145/3715005.

A COHORTS, FEATURE DEFINITIONS, AND GROUPED EVALUATION

The AIOps2025 experimental unit is a failure event with one of 11 candidate mechanisms. All 400 eligible events are used; each event enters the ten unordered pairs containing its mechanism. The primary three OOF folds keep the mechanism-by-exact-target/context cell together. The two additional grouped assignments use the same 400 events and grouping rule and passed class-support checks for every pair. In RCAEval RE1, each of Online Boutique, Sock Shop and Train Ticket has five fault labels and five injected target contexts per label. Two Online Boutique cases have injection timestamps outside the recorded telemetry interval and were excluded, leaving 123, 125 and 125 cases. Repeated fault-by-target injections remain together in three fixed grouped folds, with both labels present in every pair's train and test partitions.

Event-local summaries are computed before model fitting. AIOps2025 has 104 aligned base metrics and 20 bin means per metric. Level uses the ordered pre mean, median, population SD, min, max followed by the same five failure-interval statistics. Change uses the failure-minus-pre mean, median and SD, maximum absolute failure-bin deviation from pre mean, and normalized mean change $\Delta\bar{x}/\max(|\bar{x}_{\mathrm{pre}}|, 10^{-6})$. Temporal uses the pre/failure means and SDs, mean delta, maximum/minimum deviations, variance change, failure slope, peak magnitude and location, absolute deviation area and final-versus-peak tendency (13 values per metric). Change-Core omits only the normalized mean-change column in each metric block, leaving 416 nominal columns without recalculating any other feature.

RE1 channel names are exact metric-column identifiers from the eligible cases in each system, sorted without label information. Its pre/post segments are the recorded times before/from injection, subdivided into ten equal bins each. Availability filtering uses ≥ 0.90 finite training rows; variance filtering retains training SD $> 10^{-12}$. Training means impute missing values, and training means/SDs scale features. These steps are fitted on all training-fold events before selecting a binary pair, then applied to the corresponding held-out events. The same train-only procedure is fitted afresh under each alternate grouped fold assignment.

Table 5: Probe specifications. All use class balancing and no dataset-specific hyperparameter search; RF uses one worker.

Probe	Settings
Logistic regression	$C = 1$, `lbfgs`, `max_iter=2000`, balanced classes.
Linear SVM	$C = 1$, `dual=True`, `tol=1e-4`, `max_iter=5000`, balanced classes.
RBF SVM	$C = 1$, `gamma=scale`, `probability=False`, `cache_size=100`, balanced classes.
Random forest	100 trees, maximum depth 8, minimum leaf 2, `max_features=sqrt`, balanced subsamples, `n_jobs=1`.

The positive class is the lexicographically second fault/mechanism label. BA and macro-F1 use hard OOF predictions aggregated over all three held-out folds. In AIOps2025, AUROC additionally uses the authentic positive-class probability or SVM decision margin and is secondary. All score ties receive average ranks; exact top-set ties are resolved by canonical pair ID. Bootstrap resampling selects the same mechanism-by-context cells for all AIOps configurations within each of 10,000 draws, stratified by mechanism; RE1 draws fault-by-target cells within each fault for 10,000 draws per system. Both operate on cached OOF predictions, without model refits.

B ADDITIONAL AIOPS2025 RESULTS

Figure 2 shows every pair's rank distribution across the twelve BA representation–probe configurations. The supplement provides its complete pair-index mapping and exact ranks, all 12 representation and 18 probe BA contrasts, the corresponding macro-F1 and secondary AUROC comparisons, and the Change-Core, fit-seed and alternate-fold control results.

C RCAEval external replication

For each RE1 system and each of BA and macro-F1, the analysis compares four same-probe Level–Temporal configurations and twelve within-representation probe pairs, yielding $3 \times 2 \times (4 + 12) = 96$ contrasts. The supplementary tables give every contrast and its 2.5/97.5 percentile intervals for Spearman, Kendall tau-b, rank displacement and hardest/easiest-set overlap. The 10,000-draw paired bootstrap resamples fault-by-target cells within each fault, sharing draws across configurations and reusing cached OOF predictions. Because each system has only ten pairs, set-overlap intervals are discrete.

D OpenML cross-domain panels

The candidate list was fixed as eight named multiclass numeric tabular datasets. The six passing versions and all their classes are retained; yeast 181/v1 and page-blocks 30/v1 each fail the minimum of 30 examples in every class and are excluded before any fit. At least five classes are required. Table 6 supplies the exact versions and pair-task denominators. The selection is a prespecified non-probability set.

Table 6: OpenML source versions and exact eligible task denominators; every unordered class pair in each passing dataset is used.

Dataset	ID/version	Examples	Classes	Numeric predictors	Pairs
steel-plates-fault	40982/v3	1,941	7	27	21
gas-drift	1476/v1	13,910	6	128	15
segment	40984/v3	2,310	7	16	21
satimage	182/v1	6,430	6	36	15
pendigits	32/v1	10,992	10	16	45
texture	40499/v1	5,500	11	40	55
Total eligible					172

Each dataset uses its source ARFF row order and one shared three-fold stratified assignment (shuffle seed 20271119). Numeric predictors marked in the source metadata are filtered on all training-fold classes: finite availability at least 0.90 and observed population SD above 10^{-12}, then train-mean imputation and train-mean/SD standardization. The same fitted feature map is applied to the held-out fold before selecting each class pair. The four readouts use the specifications in Appendix A, including the primary linear SVM maximum of 5,000 iterations. Pair-level scores and ranks, all six readout contrasts under BA and same-prediction macro-F1, and hardest-30% sets are supplied as complete supplementary tables.

Within each dataset, the 10,000-replicate conditional bootstrap samples rows with replacement within every class, reusing a class's sampled indices across all pairs containing it and across all probes. Models are not refitted, so the intervals are conditional on the fixed OOF predictions and observed class samples. The supplement gives the 72 BA and macro-F1 contrast rows, their percentile intervals, exact eligibility records and source-version details.

D.1 Linear-SVM iteration-cap sensitivity

Seventeen of 516 primary OpenML Linear-SVM fits emitted an iteration-cap warning. Rerunning only these fits at 50,000 rather than 5,000 iterations produced no further warnings; two held-out hard predictions changed in one gas-drift pair. Its median BA nMARS across six contrasts changed from 0.233 to 0.229, while the other five datasets' median summaries were unchanged. The 5,000-iteration results in the main table remain primary; the higher-cap run is a numerical sensitivity check. Fit-level comparisons, affected contrasts and conditional intervals remain available in the supplement.

E WHOLE-RANKING SCALING SENSITIVITY

For two untied rankings of K tasks, the largest possible Spearman footrule distance is $F_{\max}(K) = \lfloor K^2/2 \rfloor$. Average ranks for tied scores do not exceed that bound. The main-text nMARS divides the observed mean absolute rank shift by the full span $K - 1$ available to one task. An alternative whole-ranking scale is

$$\mathrm{fMARS}(c, c') = \frac{\mathrm{MARS}(c, c')}{F_{\max}(K)/K} = \frac{\sum_p |R_p^c - R_p^{c'}|}{\lfloor K^2/2 \rfloor}. \tag{2}$$

This applies only a panel-specific positive constant to each frozen contrast; it leaves every within-panel contrast ordering unchanged. Table 7 shows the median of all four BA representation contrasts and all twelve BA probe contrasts in each panel under both scales. The nMARS and fMARS values come from the same 64 cross-panel contrast records. Under whole-ranking scaling, Online Boutique and Sock Shop retain larger probe displacement than Train Ticket. The close AIOps2025–Online Boutique comparison reverses order, so cross-panel near-ties depend on the stated scale.

Table 7: Scale sensitivity of cross-panel BA contrast medians. Each representation median includes four contrasts; each probe median includes twelve. nMARS uses the individual-task rank span and fMARS the maximum whole-ranking footrule distance.

Panel	K	Rep. nMARS	Rep. fMARS	Probe nMARS	Probe fMARS
AIOps2025	55	.140	.276	.210	.412
RE1 / Online Boutique	10	.156	.280	.217	.390
RE1 / Sock Shop	10	.122	.220	.233	.420
RE1 / Train Ticket	10	.106	.190	.094	.170

F REPRODUCIBILITY MATERIALS

The anonymous supplement contains complete pair-level rankings, configuration contrasts, conditional intervals, figure inputs, method summaries, a validation summary and numerical sensitivity results. The accompanying README indexes these materials.