Agreement Is Not Validity: A Preregistered Audit of Human-Anchored LLM-Judge Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Agreement among automatic judges is frequently reported as evidence that an LLM-based evaluator is reliable. Agreement, however, is a reproducibility property of a panel; validity concerns whether its verdicts track an independent target. We preregistered and executed an artifact reanalysis of this distinction using three public, item-level, human-referenced artifacts spanning twelve author-defined artifact–domain cells. It contrasts judge–judge Cohen's κ with the same judges' human-anchored κ and tests transport of a learned mapping from transformed reward scores to agreement with the released reference. The design uses artifact-macro aggregation, equal domain weight, and a single 10,000-draw BCa bootstrap stream. The registered frozen computation returned Gray. The H2 macro gap was 0.129636 (95% BCa CI $[0.109357, 0.151594]$): distinguishable from zero but below the registered 0.15 materiality margin. Calibration transport showed the same intermediate pattern, with macro ΔECE 0.044192 (95% BCa CI $[0.037323, 0.070487]$, Holm-adjusted $p = 0.0002$), below its 0.05 margin; ΔRisk was `GRAY-null` because only two domains had valid support. Excess agreement relative to the released human reference is measurable and artifact-dependent; it does not identify construct validity. Exploratory exclusion of PPE human ties reduces the H2 macro to 0.073393. Reference noise and label construction limit the interpretation of the gap.

1 Introduction

LLM judges score outputs, filter training data, rank systems, and provide reward signals at scale. Agreement can establish reproducibility, but shared biases, shortcuts, training exposure, or rubric mistakes can make both judges wrong. Measurement theory distinguishes reliability—measurement consistency—from validity, the warranted interpretation of what a measurement captures (Cronbach & Meehl, 1955; Messick, 1995). Inter-judge agreement supplies no independent criterion when judges share inductive biases.

Human-anchored benchmarks and recent reliability audits motivate our multi-artifact contrast (Section 2); additional evaluation background is retained in Appendix D.

We make two contributions. First, we define a same-item gap between judge–judge agreement and mean agreement with the released human reference, and estimate it without pooling items across artifacts, retaining domain and leave-one-artifact-out heterogeneity. Second, we test transport of a learned score–correctness mapping across author-defined domains, where correctness means matching that reference. These observables do not identify construct validity, shared model error, or reference noise separately. The paper is designed to remain informative under a null, a material mismatch, or the registered intermediate Gray outcome.

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.

Scope of this paper: the registered artifact arm (arm ii). All confirmatory estimates, intervals, and gates in this paper come from the three-artifact, 12-cell reanalysis. Arm i is only a four-of-five-paper codebook smoke; its full literature census was not executed, and no prevalence claim is made from it. Arm iii is outside this paper's confirmatory scope.

2 Related Work and Audit Target

Human-anchored judge evaluation. MT-Bench and G-Eval compared automatic judgments with human judgments (Zheng et al., 2023; Liu et al., 2023); answer order can also bias evaluators (Wang et al., 2024). RewardBench, PPE, and RewardBench 2 supply preference, correctness, and multi-skill comparisons (Lambert et al., 2024; Frick et al., 2024; Malik et al., 2026). Leaderboard accuracy does not establish whether panel agreement warrants a validity claim.

Agreement-centered claims. Telecom evaluation uses pairwise agreement to test its methodology (Sengupta et al., 2026); a dual-judge legal testbed uses disagreement to expose failures (Noe et al., 2026). Our target is the inference from consistency to validity. Correlated errors can make a panel look more certain without adding independent information (Kohli, 2026); structured benchmarks can expose weaknesses hidden by aggregate preference agreement (Tan et al., 2025).

Reliability, construct validity, and reference agreement. The distinction between reliability and validity is not new evidence supplied by this audit. Norman et al. (2026) evaluate judge agreement, consistency, and bias across MT-Bench, JudgeBench, and RewardBench, documenting chance-correction gaps, cross-benchmark rank instability, and high test–retest consistency coexisting with position bias; they also propose a Minimum Viable Validation Protocol. Chen et al. (2026a) assess construct validity through invariance and construct sensitivity, finding high invariance alongside low sensitivity and showing that surface-only predictors reproduce 55–67% of public labels. These studies already establish why stable judge behavior is insufficient evidence of validity. Our narrower contribution is a same-item excess-agreement contrast, $G = \kappa(a, b) - [\kappa(a, H) + \kappa(b, H)]/2$, and a test of held-out-domain transport of a learned mapping from transformed reward scores to agreement with the released reference. Norman's exact-match–κ deflation is not G; Chen's invariance/sensitivity construct-validity profile is neither G nor our calibration-transport ΔECE. Our artifact reanalysis includes RewardBench 2 (Malik et al., 2026) and uses frozen aggregation, materiality margins, and inferential gates, with disclosed non-blind amendments. The resulting Gray outcome places both macro point estimates below their registered materiality margins; selective-risk transport lacks the required domain support. These measurements characterize artifact-dependent reference agreement and its transport; they do not identify construct validity or separately identify construct mismatch, shared judge error, or reference noise. This is a limitation of our audit.

Calibration background is retained in Appendix D.

3 Preregistered Artifact-Arm Methods (Arm ii)

3.1 Eligibility and artifact roster

The preregistered feasibility floor is three qualifying H2+H3 artifacts, at least three domains, and at least 300 eligible items. These conditions are conjunctive. Eligibility criteria are specified in Appendix D.

Table 2 gives the frozen full-gate roster. Domain names and memberships are the artifact authors' task families. PPE memberships overlap; when one PPE domain is held out, the same item is removed from every other PPE training slice to prevent leakage. "Overall" is reporting-only. A deterministic verbosity heuristic is retained as one PPE baseline but is not described as a learned reward model.

Table 1: Novelty contrast: nearby evaluations and the estimands they answer.

Audit family	Primary target	Difference from this paper
Human-agreement benchmarks (Zheng et al., 2023; Liu et al., 2023; Frick et al., 2024)	Judge performance against human preferences	Do not estimate a judge–judge minus same-item judge–human gap across artifacts.
Single-domain agreement audit (Mao, 2026)	Automatic-panel agreement and human anchoring in one domain	Does not estimate a frozen multi-artifact macro or roster sensitivity.
Large-scale judge audit (Norman et al., 2026)	Judge–human agreement, test–retest consistency, and position/verbosity bias; minimum validation protocol	Exact-match–κ deflation differs from our same-item excess-agreement estimand G and reference-agreement calibration transport.
Construct-validity profile (Chen et al., 2026a)	Invariance, construct sensitivity, and surface-only prediction of 55–67% of public labels	Invariance/sensitivity is neither G nor ΔECE; our audit does not identify construct validity.
Correlated-error and failure audits (Kohli, 2026; Song et al., 2026)	Shared errors and surface-heuristic consensus	Do not jointly gate the same-item gap and held-out-domain calibration transport.
This paper	Artifact–domain gap G and transported ΔECE/ΔRisk	Preregistered weighting, margins, uncertainty, and artifact-deletion map.

Table 2: Frozen artifact roster. Counts and transformations are fixed by the artifact manifests and Amendments A5–A6, not chosen from effects.

Artifact	Human reference	Automatic configurations	Analytic domains	Items
RewardBench	curated chosen response	two Matter configurations	Chat; Chat Hard; Safety	1,145
PPE Human Preference V1	Arena user vote	14 traces (13 learned, 1 heuristic)	six overlapping families	16,038
RewardBench 2	published per-item verification	two staged reward models	Math; Safety; Precise IF	793

Artifact-specific eligibility, label construction, and score transformations are detailed in Appendix D.

3.2 Orientation, agreement, and the H2 estimand

The orientation correction and the hash h_N used for presentation and folds are specified in Appendix D.

All primary agreement statistics are unweighted Cohen's κ (Cohen, 1960). For automatic judges a, b, human reference H, artifact j, and domain d, all three labels must exist on the same complete cases and at least 50 common items are required. The pair contrast is

$$G_{abjd} = \kappa(a, b) - \frac{[\kappa(a, H) + \kappa(b, H)]}{2}. \tag{1}$$

Thus G_{abjd} subtracts the arithmetic mean of the two human-anchored kappas from judge–judge kappa. Undefined κ from degenerate marginals excludes the pair. Within an artifact–domain cell, H2 takes the median across eligible judge pairs. A domain estimate is the arithmetic mean across artifacts carrying that domain, and the overall estimate is the unweighted mean across the twelve distinct artifact–domain cells. Items are never pooled across artifacts. In this frozen roster, each analysis-domain key identifies one artifact–domain cell;

similarly named task families in different artifacts remain separate, and the domain counts in the registered gates use these keys.

This implementation gives every cell equal weight: PPE's six cells contribute $6/12 = 0.50$ of the macro, and RewardBench and RewardBench 2 contribute $3/12 = 0.25$ each. Thus no artifact exceeds the registered 50% ceiling, but PPE sits exactly at it. The material H2 margin is $|G| = 0.15$; either sign is publishable. Positive values mean automatic agreement exceeds mean agreement with the released human reference. We call G *excess agreement relative to the released human reference*; it is not an identified validity deficit. Even conditionally independent judges can have positive G when the reference is noisier than the judges (Section E.3).

3.3 Calibration transport and H3

H3 tests transport of a learned score–correctness mapping, with correctness defined relative to the released reference. Its input is a transformed reward score, not an elicited confidence statement. For each judge-item, $y = 1$ iff its emitted verdict equals H. For each judge within each artifact, the held-out-domain calibrator g_{-d} is fit on that judge's eligible observations from the other domains of the same artifact, with each nonempty training domain receiving equal total weight. For PPE, all fixed held-out item IDs are excluded from every other training slice. The comparator g_d is five-fold out-of-fold isotonic calibration for the same judge within d. Fold assignment and fit floors are specified in Appendix D.

ECE uses ten fixed equal-width bins. At calibrated confidence q,

$$\mathrm{ECE} = \sum_{b=1}^{10} \frac{n_b}{n} \left| \overline{q}_b - \overline{y}_b \right|. \tag{2}$$

Here n is the number of evaluated judge–item observations for the calibration evaluation, n_b is the count in bin b, and $\overline{q}_b$ and $\overline{y}_b$ are that bin's mean calibrated probability and mean reference-agreement indicator. Empty bins contribute zero; bins are left-closed and right-open, except that the final bin includes $q = 1$. Selective risk is the error fraction among items with $q \geq 0.80$ and is valid only with at least 25 accepted items and 20% coverage. A deterministic top-coverage sensitivity retains $\lceil 0.8n \rceil$ items. Within each artifact, eligible judges are macro-averaged; artifact contrasts are then averaged without pooling items. H3 reports

$$\Delta\mathrm{ECE}_d = \mathrm{ECE}_{-d} - \mathrm{ECE}_d, \qquad \Delta\mathrm{Risk}_d = \mathrm{Risk}_{-d} - \mathrm{Risk}_d.$$

At least three valid held-out domains are required per measure. Material transport margins are 0.05 ECE and 0.10 risk; positive values indicate worse transport.

3.4 Bootstrap, multiplicity, and registered gates

H2 and H3 use 10,000 paired BCa draws (Efron, 1987) from one NumPy `PCG64(20260831)` stream consumed in the fixed order H2 macro, H3 ECE macro, H2 domain intervals, and H3 ECE domain intervals. No ΔRisk bootstrap stage was run because fewer than three domains had valid support; it consumed no random draws, and its unsupported test was assigned $p = 1$ for the two-member Holm adjustment. Resampling, overlap handling, and BCa mechanics are detailed in Appendix D.

H3's two paired-bootstrap p-values, ΔECE and ΔRisk, form one Holm family (Holm, 1979). H2 is gated by its two-sided BCa interval and magnitude, not a separate p-value. Domain-specific intervals and leave-one-artifact-out estimates are mandatory diagnostics, not additional confirmatory families.

The CPU decision is Proceed if H2 has $|G| \geq 0.15$, its 95% interval excludes zero, and at least two domains have the same sign and $|G_d| \geq 0.15$; or if an H3 macro clears its margin, the corresponding Holm-adjusted $p < .05$ and interval exclude zero, and at least two held-out domains clear the same-direction margin. The registered KILL / honest-null rule is: H2 has $|\text{macro gap}| < 0.05$, CI wholly inside $[-0.10, +0.10]$, and no domain reaches $|0.15|$; H3 has $|\text{macro } \Delta\text{ECE}| < 0.02$ with CI inside $[-0.05, +0.05]$, $|\text{macro } \Delta\text{Risk}| < 0.05$

Table 3: Frozen H2 excess-agreement gaps relative to the human reference by artifact–domain cell. Intervals are 95% BCa intervals from the registered bootstrap. RB denotes RewardBench; PPE denotes PPE Human Preference V1; RB2 denotes RewardBench 2.

Artifact–domain	Gap	95% CI
RB / Chat	−0.091839	[−0.148968, −0.025059]
RB / Chat Hard	0.048205	[−0.101841, 0.218207]
PPE / Coding	0.260815	[0.242291, 0.280237]
PPE / Easy	0.189147	[0.172823, 0.208166]
PPE / Hard	0.259077	[0.243814, 0.273302]
PPE / Instruction Following	0.242316	[0.225948, 0.258939]
PPE / Math	0.267993	[0.253314, 0.290971]
RB / Safety	0.266672	[0.175470, 0.363884]
PPE / Similar Responses	0.227556	[0.214338, 0.237330]
RB2 / Math	−0.043344	[−0.125454, 0.048470]
RB2 / Precise IF	0.105190	[−0.009722, 0.228249]
RB2 / Safety	−0.176152	[−0.205867, −0.143873]

with CI inside $[-0.10, +0.10]$, and no domain reaches either material threshold. Every other integrity-passing result is GRAY. The optional capability-gradient arm cannot rescue a failed CPU gate. Figure 2 (Appendix D) diagrams the estimands and gates.

4 Execution, Verification, and Amendments

The completed report has inference status `REGISTERED-FROZEN`, integrity pass, and SHA-256 prefix `cdc6b959c652`.

A5 and A6 are explicitly non-blind and are the main threats to a fully outcome-blind confirmatory interpretation. Full smoke, mirror-refutation, checkpoint, amendment, and deviation records are retained in Appendix C; the amendment log is Table 5.

5 Registered Artifact-Arm Results

The frozen report passed integrity and returned the following registered gate sentence (retained verbatim as an archival record; its legacy word "validity" denotes reference agreement, not identified construct validity):

> "Decision GRAY: H2 judge-agreement/validity macro gap 0.1296 (CI [0.1094, 0.1516]) — real, CI excludes 0, but the estimate is below the registered 0.15 materiality line; H3 dECE 0.0442 (CI [0.0373, 0.0705], Holm p 2e-4) — real, below the 0.05 margin; dRisk GRAY-null (2 valid domains)."

The archived names `dECE` and `dRisk` denote ΔECE and ΔRisk, respectively.

5.1 Full artifact roster and H2 artifact–domain results

The frozen report passes integrity and the Section 3.1 portfolio floor with 3 artifacts, 12 artifact–domain cells, and 17,976 unique eligible items. Table 3 gives the frozen H2 artifact–domain estimates and intervals.

5.2 Gray outcome: heterogeneous, below-margin macros

Automatic-judge agreement and agreement with the released human reference differ, but the registered materiality conjunction is not met. The H2 macro gap is 0.129636 with 95% BCa interval $[0.109357, 0.151594]$. The interval excludes zero, and multiple PPE domains have positive gaps above the 0.15 domain margin, but the macro point estimate remains below 0.15. RewardBench and RewardBench 2 prevent a simple "all artifacts agree" story: the frozen report includes negative cells for Chat and RewardBench 2 Math and Safety, including

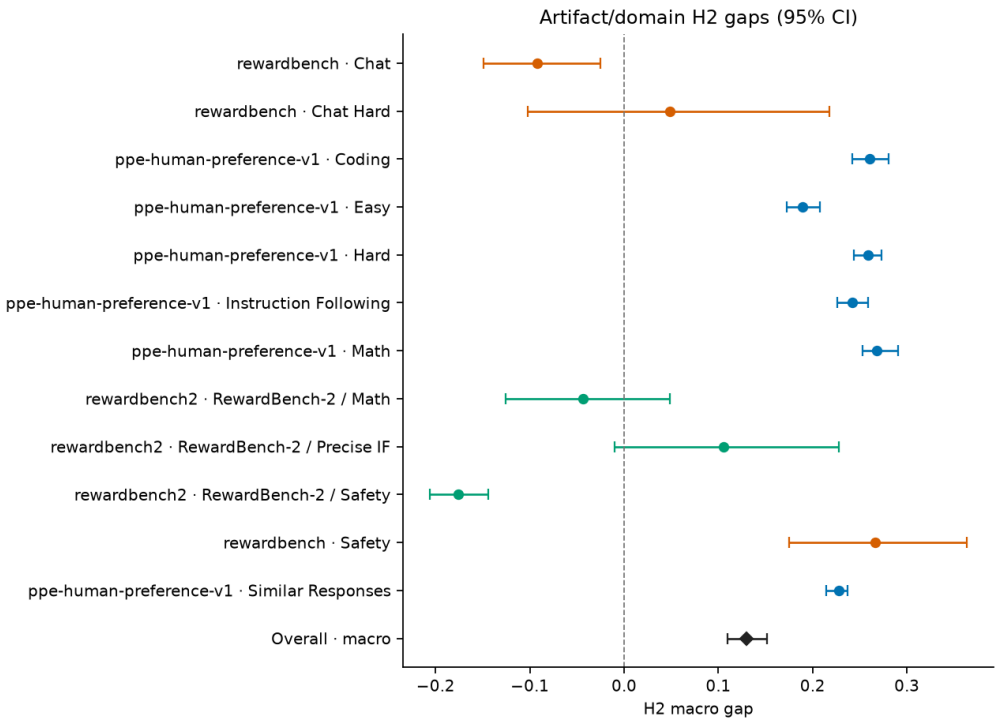

Figure 1: Frozen H2 artifact–domain estimates and the registered macro.

−0.176152 for RewardBench 2 Safety. Negative cell gaps indicate lower judge–judge kappa than mean judge–reference kappa; the illustrative reference-noise construction in Section E.3 also permits negative gaps when its reference is more accurate than its judges, but does not identify the mechanism in these artifacts. The observed excess agreement is strongly artifact-dependent, and the preregistered H2 effect is below the material macro margin. This comparison alone establishes neither a construct-validity failure nor the adequacy of agreement as a proxy for validity.

Transport of the learned score–correctness mapping has the same intermediate shape. The macro ΔECE is 0.044192, with 95% BCa interval $[0.037323, 0.070487]$ and Holm-adjusted $p = 0.0002$; the point estimate remains below the 0.05 materiality margin. Some held-out domains exceed that margin, but statistical detectability does not substitute for the registered effect-size criterion. ΔRisk is `GRAY-null` because only 2 domains retain registered selective-risk support, fewer than the required three. Negative cell-level ΔECE values mean that the outside-domain calibrator has lower ECE than the within-domain out-of-fold comparator on the evaluated items; each negative point estimate in Table 4 has an interval containing zero, so these cells do not establish improved transport.

ΔRisk was valid only for Chat (−0.019117) and RewardBench 2 Safety (−0.038950), so its macro is `GRAY-null`. Artifact–domain estimates and intervals appear in Table 4.

The domain-effect panel is Figure 3 in Appendix E.

5.3 Artifact dependence and sensitivity

Across 2,954 complete PPE Coding items, Athene-RM-70B and Athene-RM-8B agree on 87.47% of labels, versus 41.47% and 40.28% agreement with the human vote, respectively. The resulting pair gap is 0.605931, even though the Coding cell median across all 91 pairs is 0.260815. This is a matrix-level worked case, not a reconstructed individual item. The full worked-matrix record is retained in Appendix E.8.

Table 4: Frozen H3 artifact–domain ΔECE estimates and 95% BCa intervals. ΔRisk was valid only for Chat (−0.019117) and RewardBench 2 Safety (−0.038950), so its macro is `GRAY-null`. RB denotes RewardBench; PPE denotes PPE Human Preference V1; RB2 denotes RewardBench 2.

Artifact–domain	ΔECE	95% CI
RB / Chat	0.233004	[0.197471, 0.388434]
RB / Chat Hard	−0.032368	[−0.106644, 0.039699]
PPE / Coding	0.008606	[0.003986, 0.022275]
PPE / Easy	0.016004	[0.008818, 0.033312]
PPE / Hard	0.008563	[0.002448, 0.018767]
PPE / Instruction Following	0.010372	[0.004410, 0.021394]
PPE / Math	0.063215	[0.053483, 0.085065]
RB / Safety	0.172862	[0.121163, 0.216390]
PPE / Similar Responses	0.026390	[0.017851, 0.039508]
RB2 / Math	−0.002989	[−0.061044, 0.074124]
RB2 / Precise IF	−0.008140	[−0.052084, 0.102435]
RB2 / Safety	0.034789	[−0.018182, 0.107582]

Table 10 (Appendix E) reaverages the frozen H2 domain estimates after removing all cells from one artifact. These deterministic diagnostics have no separate bootstrap interval or gate. Removing PPE reduces the mean nearly to zero; removing RewardBench gives 0.148066, about 0.002 below the 0.15 margin; and removing RewardBench 2 gives 0.185549, above that margin. The three cases neutrally summarize the roster sensitivity of the macro.

These results frame the contribution as an artifact-heterogeneity measurement. Adding RewardBench 2 changes the mix of human-reference construction, label cardinality, and judge families; its domain estimates are not interchangeable with PPE's Arena-preference estimates. The data reject a one-number account of "judge reliability," while the gate rejects the stronger claim that the average mismatch clears the preregistered practical margin. The appropriate headline is: the size and direction of excess agreement relative to the released reference depend on the artifact. Construct validity remains unidentified by this contrast alone.

The artifact-weight and heterogeneity panel is Figure 4 in Appendix E.

These post-review analyses use fresh aggregation code and retained raw scores; they have no confirmatory intervals, multiplicity decisions, or gate status. The registered estimates and GRAY verdicts remain unchanged. Human ties occupy 36.8–46.9% of the six PPE domain slices, whereas the registered automatic rule requires exact score equality. Excluding human-tie PPE items, while retaining all other artifacts, all judge pairs, pairwise complete cases, the 50-item floor, and twelve equal cell weights, gives $G = 0.073393$ in the independent raw-artifact recount. The macro falls by 43.39%: that fraction of the original estimate is sensitive to removing human ties. This is not a causal allocation to adapter error: deletion changes the item population, class marginals, and chance correction. The remaining gap still measures reference disagreement on decisive human preferences, not a validity deficit.

Full adapter, reference-noise, calibrator, operating-point, artifact-deletion, weighting, and worked-matrix diagnostics are retained in Appendix E.

Table 9 (Appendix E) reports retained operating points; the protocol's full risk–coverage curve requirement was not fulfilled by the archived report.

6 Discussion

Agreement is evidence, but evidence for a narrower claim. Inter-judge agreement measures panel consistency; construct validity requires a warranted link between reference and construct. Reference errors can create the same positive gap as shared model error. The present artifacts cannot distinguish construct mismatch, shared model error, and reference noise. Nominal label support also does not equate human and automatic tie rules.

Artifact heterogeneity is a result, not nuisance alone. The artifacts differ in reference construction, judges, score semantics, and domains (Table 2). Equal artifact–domain weighting targets the average author-defined cell, not the average item, making domain and leave-one-artifact-out results central to interpretation.

Practical validation claims. Judge-validation claims should identify the construct, compare the deployed judge with an independent same-item reference, disclose marginal prevalences and confusion patterns, and test operational confidence calibration. Multiple judges probe sensitivity but cannot establish reference reliability or its relationship to the construct.

What the gate does not say. A GRAY outcome is not an equivalence finding, and a PROCEED outcome is not a universal indictment of model evaluation. Both are scoped to the frozen artifacts, domains, judge configurations, and weights. The strict honest-null branch would require substantially tighter H2 and H3 null conditions than merely missing a materiality threshold. Even that branch would bound the registered agreement contrast and transport errors, without establishing construct validity or a reliable reference.

7 LIMITATIONS

Judge-family coverage. The roster is broad inside PPE but uneven across artifacts. RewardBench's two configurations share a base model. RewardBench 2 contributes only the two staged minimum traces, BTRM-Qwen2-7B and RAMO-Llama3.1-8B. PPE has thirteen learned reward models plus a deterministic verbosity baseline, but availability is confounded with artifact. We cannot separate judge family, scale, training data, and artifact provenance causally. The optional arm-iii capability gradient is descriptive and cannot repair that confounding.

Human-reference quality. Human anchoring is not error-free ground truth. PPE contains single user votes and explicit ties; RewardBench's included subsets mix manual and curated construction. RewardBench 2 admission relies on published statements of per-item verification for Math, Safety, and Precise IF; raw annotator identity, votes, and adjudication are not present in the staged artifact, and no author attestation was sought. Ambiguous raw-human pluralities would be excluded, but that rule cannot recover annotation metadata that was never released. There is no identified reference-reliability ceiling here. The binary reference-noise construction in Section E.3 demonstrates non-identification; it is not an estimate of reliability for these artifacts.

Domain weighting. The primary macro gives equal weight to twelve artifact–domain cells. Because PPE supplies six, it receives exactly half of the total weight while each other artifact receives one quarter. PPE categories overlap, so one underlying item can inform several cell estimates even though held-out leakage is removed from calibrator training. Because each of the twelve cells contains exactly one artifact, every CI is conditional on the frozen roster; the leave-one-out results show that roster changes move the macro from 0.018122 near the zero line to 0.185549 above the 0.15 line. Per-cell item resampling also does not preserve positive cross-cell covariance induced by PPE overlap. The BCa acceleration also uses deletions confined to a single artifact–domain cell rather than joint deletion of all memberships of an overlapping PPE item, so it does not restore the shared-item covariance omitted by the bootstrap. Alternative item-pooled or artifact-equal summaries answer different questions; their deterministic point estimates are reported in Table 11 (Appendix E) and remain non-registered secondary analyses.

Calibration support. Fixed fit floors protect isotonic calibration from extreme low-support fits but can make selective-risk contrasts undefined. ECE depends on ten fixed bins and does not capture every calibration defect. The 0.80 selective threshold is one operating point. The retained operating-point records and deterministic top-coverage sensitivity are reported in Table 9 (Appendix E); they show that support, more than the direction of the top-coverage contrast, determines the registered ΔRisk status.

Non-blind amendments and compute verification. A5 followed observation of degenerate RewardBench kappas, and A6 followed the two-artifact midscale look. Both are transparently frozen before the full gate, but neither is outcome-blind from study inception. Separately, refuting the Numba mirror demonstrates that implementation equivalence should be tested at the inferential endpoints, not inferred from matching point estimates. The pure-Python run and checkpoint discipline protect the registered calculation, not against errors shared by protocol and implementation.

Temporal and publication scope. The motivating audit with four eligible of five screened is only a codebook smoke and cannot support a literature-prevalence claim. Artifact availability selects for transparent projects and may underrepresent proprietary evaluation practice. New papers, later revisions, and unreleased judge traces can change the descriptive landscape, while the frozen artifact estimand remains unchanged.

8 Conclusion

Agreement among models can establish consistency; it cannot establish what the models are consistently measuring. Across three public artifacts, the frozen same-item analysis found a nonzero H2 macro gap and nonzero calibration transport error, but both point estimates remained below their registered materiality margins. The resulting Gray verdict is a heterogeneous measurement map, not an equivalence claim: removing RewardBench 2 would have moved the H2 diagnostic above 0.15, while its own cells include two negative gaps. Exploratory analyses show sensitivity to tie construction, reference noise, and calibrator choice. Consensus should be reported as consensus; validity requires additional evidence about the construct and reference at the artifact and domain where the judge will be used.

AI use statement

We used large language models as writing assistants to improve clarity and polish text, retrieve literature and assist with reference verification, support research ideation and execution, and draft manuscript sections. Under author direction, AI coding agents implemented and executed experiments according to preregistered protocols; numerical results were checked through independent recounts. The authors directed this work, reviewed the outputs, and retained responsibility for verification, interpretation, and all research decisions. All three authors reviewed the AI-assisted manuscript and take full responsibility for its final content, including AI-assisted text, claims, results, and artifacts.

Reproducibility statement

The public artifact is available at `https://anonymous.4open.science/r/agreement-audit-2027` (anonymized mirror during review) at release commit `7ef431fbbd25db8496391f428e3251b125919e4c`. The archived computation is pinned to Git commit `52a206b41ac04245769ee4b595c74c83a49da8f8`. From the repository root, `pilots/R3/p1/` contains ingestion and gate code plus frozen artifact manifests, `pilots/R3/p1/full/` contains the checkpoint, run log, and frozen report, and `paper/R3-P1/` contains manuscript source, figures, and rendered text.

The exact fresh registered invocation is:

```
python3 pilots/R3/p1/run_full_gate.py --backend frozen \
  --output /tmp/armii-full-report-reproduced.json
```

The provenance-preserving checkpoint invocation is:

```
python3 pilots/R3/p1/run_full_gate.py --backend frozen \
  --resume-from pilots/R3/p1/full/checkpoint-45h.json \
  --output /tmp/armii-full-report-reproduced.json
```

The registered backend is intentionally explicit because the default Numba backend is preview-only. The archived SHA-256 references are:

Frozen report	`cdc6b959c6522888059cab5c17857b2a58bfe96b2703f6ab472711121ef2ec79`
Primary freeze manifest	`7f2a2d01465e4744caf7c0f8ee59a88be050aec07758822be5970bdb78b8fe65`
RewardBench 2 manifest	`52123c63774dc964a0e40d5d1e9dfe8e9125c08a0841c802c552caa7296afd28`

The report's own integrity block verifies every underlying artifact file against those manifests. The deterministic secondary H2 reweightings in Appendix E.7 are reproduced with `python3 paper/R3-P1/sensitivity_from_frozen.py`.

References

Jianlin Chen, Wenhui Chen, Ziyao Lin, and Chi Man Vong. A judge should know what changed: Construct validity for LLM-as-a-judge evaluation, 2026a. arXiv:2608.24419; v1 submitted 25 August 2026.

Wenbo Chen, Veena Padmanabhan, Tootiya Giyahchi, Elaine Wong, and Leman Akoglu. Rethinking evaluation for LLM hallucination detection: A desiderata, a new RAG-based benchmark, new insights, 2026b. arXiv:2605.11330.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 8359–8388, 2024. arXiv:2403.04132.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF, 2024. arXiv:2410.14872.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

Shohei Hisada, Endo Sunao, Himi Yamato, Shoko Wakamiya, and Eiji Aramaki. Filling in the clinical gaps in benchmark: Case for HealthBench for the japanese medical system, 2025. arXiv:2509.17444.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels, 2026. arXiv:2605.29800.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Annenberg School for Communication, University of Pennsylvania, 2011. URL `https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf`. Dated January 25, 2011; literature updated September 13, 2013.

Dhruv Kulkarni and Sai Manoj Pudukotai Dinkarrao. MultModLM: A multi-modal benchmark for large-language model based hardware schematic generation, 2026. arXiv:2606.27666.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling, 2024. arXiv:2403.13787.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. arXiv:2303.16634.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, 2026. arXiv:2506.01937.

Hui Mao. Historical backtesting for scientific question discovery: A protocol and astronomy pilot, 2026. arXiv:2608.16795.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749, 1995.

Su Myat Noe, Ha Thanh Nguyen, May Myo Zin, and Ken Satoh. Beyond accuracy: A dual-judge evaluation protocol for vision-language models in legally grounded tasks, 2026. arXiv:2608.24258; AI for Law Workshop at ICML 2026.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias, 2026. arXiv:2606.19544.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. arXiv:1906.02530.

Rishiraj Sengupta, Sotiris Chatzimiltis, Mohammad Shojafar, and Xiatian Zhu. Free-text evaluation of LLMs for 5G domain knowledge and fault analysis using LLM-as-judge, 2026. arXiv:2608.21021; accepted at IEEE CSCN.

Mingyang Song, Mao Zheng, and Chenning Xu. Beyond the illusion of consensus: From surface heuristics to knowledge-grounded evaluation in LLM-as-a-judge, 2026. arXiv:2603.11027.

Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, and Ziyan Wu. MedGRPO: Multi-task reinforcement learning for heterogeneous medical video understanding, 2025. arXiv:2512.06581.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025. arXiv:2410.12784.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 9440–9450, 2024. arXiv:2305.17926.

Jing Yu, Yibo Zhao, Jiaming Zhang, and Xiang Li. LazyMem: Retrieve broadly, construct selectively for efficient long-term agent memory, 2026. arXiv:2607.22690.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023. arXiv:2306.05685.

A Machine-Readable Confusion and Marginal Records

The frozen report retains the complete confusion matrix and both marginal prevalence vectors for every eligible judge–judge and judge–human comparison. They are nested under each artifact–domain cell in `h2.details`, inside the corresponding `pair_kappas` record. This preserves the exact integer cells and floating-point marginals for all 91 PPE judge pairs and the two one-pair artifacts without compressing hundreds of model-specific matrices into an unreadable paper table. The same records give the common-item count, kappa, label support, and any excluded pair; the frozen report records no excluded H2 judge pairs.

As registered sensitivity checks, the same pairwise contrast and cell-median aggregation applied to raw percent agreement gives a macro gap of 0.146223; nominal Krippendorff's α (Krippendorff, 2011), recomputed from the stored confusion matrices, gives 0.167713. Both preserve the primary contrast's positive direction. These sensitivities diagnose dependence on the chance-correction convention; they do not replace the registered Cohen's-κ gate.

B Frozen-Report Verification

The report records 10,000 draws from the single registered `PCG64(20260831)` stream, zero rejected H2 macro draws, zero rejected H3 ΔECE macro draws, completed domain-specific intervals, and mechanical decision `GRAY`. The paper tables retain six decimals from that report; the quoted gate sentence uses the registered rounded reporting form.

C Execution, Amendments, and Verification Provenance

C.1 Execution summary

The smoke and staged changes are detailed in Section C.2.

The frozen-inference verification is detailed in Section C.3. Checkpoint resume did not alter evaluation order, random draws, estimators, thresholds, or data.

C.2 Motivating smoke audit and staged changes

The registered codebook smoke had four eligible of five screened by the frozen hash rule. The medical-video benchmark (Su et al., 2025), hallucination benchmark (Chen et al., 2026b), and agent-memory study (Yu et al., 2026) were adjudicated `HUMAN-ANCHORED`; the Japanese HealthBench study (Hisada et al., 2025) was `OTHER-VALIDATION`. The hardware-schematic benchmark (Kulkarni & Pudukotai Dinkarrao, 2026) was ineligible because its judged artifact was an image rather than natural-language text. The labels for the four eligible of five screened are *smoke-scale evidence only*, not a census, and support neither a prevalence estimate nor a Wilson interval. The protocol's full literature census remains future work. The smoke disagreement exposed two ambiguities: cited human validation of someone else's judge is not a human anchor for the paper's own judge, and expert rubric authorship is not an item-level human comparison. Amendment A4 pinned both rules before midscale coding.

The artifact-arm smoke verified parsing and hand fixtures on RewardBench but exposed the constant-orientation κ defect. A5 records the resulting non-blind derandomization. Midscale then included RewardBench and PPE. It passed data and estimator integrity but necessarily failed the conjunctive feasibility floor: two artifacts cannot support the registered three-artifact inference, and the implied PPE artifact weight exceeded the 50% ceiling. The effects were viewed, so A6's subsequent admission of RewardBench 2 is explicitly non-blind. It freezes the three admissible human-verified subsets, two judge traces, four-class score interpretation, and h_4 rotation before any three-artifact gate run.

C.3 Fast-mirror refutation and frozen inference

The midscale computation used a Numba-accelerated mirror intended to reproduce the registered Python estimator. An independent frozen-path recount compared 105 outputs. Every point estimate matched exactly, but jackknife acceleration constants and the BCa endpoints derived from them differed; the maximum absolute discrepancy was 2.5×10^{-4}. The exact-equality verification criterion therefore returned `MIRROR-REFUTED`. Although this difference was too small to move a then-visible gate condition, the mirror was demoted to preview-only. No interval or p-value from that backend is confirmatory.

The registered full inference called the frozen high-level pure-Python estimators for every bootstrap draw, rejected-draw replacement, jackknife deletion, and BCa calculation. Point-estimate preflight matched the report path exactly. The slower path is not presented as superior numerically; it was used because it is the implementation frozen by Amendment A2 and the mirror failed its equality contract. The completed report has inference status `REGISTERED-FROZEN`, integrity pass, and SHA-256 prefix `cdc6b959c652`.

C.4 Checkpoint and resume provenance

The first frozen run was terminated by a session restart after approximately 45 hours. Its atomic checkpoint preserved the completed H2-macro and H3-ΔECE-macro stages (roughly 40 hours); the in-flight H2 domain stage was discarded. The checkpoint-resume path validated artifact identity and point estimates, restored the registered random stream by exact sampling replay through the two completed stages, and then skipped them. The serialized H2-macro and H3-macro payloads in the 45-hour checkpoint and the resumed report have identical SHA-256 digests when canonicalized together. Thus the completed stages were reused verbatim, not reconstructed from rounded summaries; only unfinished downstream

Table 5: Protocol amendment log. No amendment changes a materiality margin.

ID	Timing	Pin	Blindness
A1	pre-data	Exact h_5 folds; no label-based fold rebalancing	blind
A2	pre-data	strict-less BCa bias count, linear quantiles, one PCG64 stream	blind
A3	pre-data	NaN-κ exclusion, ceil top coverage, fold-coverage recording	blind
A4	post-smoke, pre-midscale	text-output eligibility and own-judge human-anchor rules	smoke seen
A5	post-smoke, pre-midscale	h_2 orientation derandomization for constant references	non-blind
A6	post-midscale, pre-full	RewardBench 2 provenance, subsets, judges, logits, and h_4 rotation	non-blind

stages were recomputed. This 45-hour checkpoint-resume provenance is disclosed because the run log records every start, completion, checkpoint, RNG restore, and skip.

C.5 Amendments and deviations

Table 5 summarizes all six amendments. A1–A3 were written before artifact coding. A4 followed the classification smoke with four eligible of five screened but preceded midscale. A5 and A6 are explicitly non-blind and are the main threats to a fully outcome-blind confirmatory interpretation.

Protocol Deviation note D1 (2026-09-07) records the arm-i scope reduction: arm i was executed only as the smoke with four eligible of five screened, while the registered full census remains scheduled; this paper therefore carries arm ii as its registered contribution and uses the $n = 4$ arm-i labels only as motivating evidence. D1 is a disclosed scope deviation, not an amendment: neither an arm-i procedure nor estimand changed. Separately, rejecting the Numba mirror was an implementation refutation, not an inferential deviation; the registered frozen path remained the sole source of confirmatory inference. Checkpoint resume did not alter evaluation order, random draws, estimators, thresholds, or data. Arm iii remains a conditional descriptive extension and is not required for this paper.

D Artifact Eligibility and Estimator Mechanics

Evaluation background. The distinction matters in practice. MT-Bench popularized strong LLM judges by measuring agreement with human preferences as well as operational consistency (Zheng et al., 2023); G-Eval similarly compared automatic scores with human judgments (Liu et al., 2023). Later judge and reward-model benchmarks made human preference agreement a central target (Lambert et al., 2024; Frick et al., 2024; Tan et al., 2025), while Chatbot Arena established the crowdsourced preference source underlying PPE (Chiang et al., 2024). But high automatic–automatic agreement can coexist with much weaker agreement to humans on the same items, as a recent single-domain astronomy study illustrates (Mao, 2026). Norman et al. document chance-correction gaps, cross-benchmark rank instability, and high repeatability coexisting with position bias (Norman et al., 2026). Chen et al. assess construct validity through invariance and construct sensitivity (Chen et al., 2026a); other recent audits document correlated panel errors and surface-heuristic agreement (Kohli, 2026; Song et al., 2026). These results motivate a direct multi-artifact estimand rather than another judge leaderboard.

Calibration and transport. Calibration of a score-derived prediction is separate from hard-label agreement. Our inputs are softmax-transformed reward scores, not judges' stated confidence. Expected calibration error (ECE) summarizes the gap between predicted probability and observed reference agreement (Guo et al., 2017), and calibration is known to degrade under dataset shift (Ovadia et al., 2019). We focus on transport: whether a calibrator learned outside a held-out domain is worse than an out-of-fold calibrator fit inside it. This comparison makes the domain shift explicit and avoids equating a good pooled reliability diagram with deployable cross-domain confidence.

D.1 Eligibility criteria and feasibility floors

An artifact must be public and redistributable or checksum-referenceable, retain at least 50 eligible items, contain an independent item-level human label, include outputs from at least two automatic judges on the same items, and expose an admissible calibration input (called "confidence" in the registration). This is a released probability, a two-class logit/score pair, or at least five repeated votes; bounded ordinal ratings are not reinterpreted as probabilities. The preregistered feasibility floor is three qualifying H2+H3 artifacts, at least three domains, and at least 300 eligible items. These conditions are conjunctive.

For the repeated-vote route, the implemented helper requires at least five votes, emits a plurality label, and uses its vote count divided by the total number of votes as the score. Count ties are resolved among lexicographically sorted tied labels using h_N of the comma-joined artifact, item, and judge IDs; the tied maximum frequency remains the score. None of the retained artifacts uses this route.

D.2 Artifact-specific eligibility and score construction

RewardBench. We retain twelve source subsets conservatively documented as manually confirmed or curated, yielding 1,145 pairs in Chat, Chat Hard, and Safety (Lambert et al., 2024): `alpacaeval-easy`, `alpacaeval-length`, `alpacaeval-hard`, `mt-bench-easy`, `mt-bench-med`, `mt-bench-hard`, `llmbar-natural`, `llmbar-adver-manual`, `xstest-should-respond`, `xstest-should-refuse`, `refusals-dangerous`, and `refusals-offensive`. The two released Matter configurations share a base model but differ in scoring configuration. Each supplies chosen and rejected scores; the emitted label is the larger score and confidence is its float64 two-way softmax probability.

PPE. PPE contributes 16,038 Chatbot Arena battles and fourteen aligned score traces (Frick et al., 2024). The user vote maps to A, B, or an explicit Tie. Each reward trace supplies two response scores; for unequal scores, H3 uses the emitted label's float64 two-way softmax probability, $s = \exp(r_{\max}-m)/[\exp(r_A-m)+\exp(r_B-m)]$, where $m = \max(r_A, r_B)$. Equal automatic scores emit Tie for H2 but are missing for H3 because a two-score vector gives no probability to a third class. Error sentinels and malformed scores are missing, never

ties. The six analytic domains are Hard, Easy, Instruction Following, Coding, Math, and Similar Responses.

RewardBench 2. Amendment A6 admits exactly Math (183 items), Safety (450), and Precise IF (160): the paper states that every example or sample in these subsets was manually verified (Malik et al., 2026). Factuality, Focus, and Ties are excluded. The release contains four candidate completions per retained item. For each of two staged reward-model traces, the four scores are logits; the emitted nominal label is the argmax in $\{A, B, C, D\}$ and confidence is the float64 four-way softmax probability of that class. Exact maxima ties are H2 ties and H3 missing. The human-verification provenance is paper-stated rather than independently audited, a limitation we retain explicitly.

D.3 Orientation correction

Constant answer position makes unweighted κ degenerate: if every human reference is A, expected and observed marginal agreement collapse so judge–human κ can be zero regardless of accuracy. Amendment A5 therefore derandomizes RewardBench presentation by

$$h_N(x) = \text{int}(\text{SHA256}(x)[0{:}8], \text{big}) \bmod N.$$

The input x is encoded as UTF-8; [0:8] selects the first eight raw digest bytes, interpreted as an unsigned big-endian integer. RewardBench uses the released item's ID string. RewardBench 2 uses the compact JSON array of the string-valued subset and ID, serialized with `ensure_ascii=False` and `separators=(",", ":")`. For RewardBench, h_2 swaps the response positions, human label, judge labels, and score slots together. Amendment A6 applies the same information-preserving operation to RewardBench 2 as an h_4 rotation of the stable `(subset,id)` key. PPE already has naturally varying $A/B/\text{Tie}$ human labels and is not rotated.

D.4 Calibration fold mechanics

Folds apply the same h_5 to the item-ID string: the released ID for RewardBench, the string-valued `question_id` for PPE, and the compact JSON key above for RewardBench 2. There is no outcome-based rebalancing. A fit is suppressed when a training fold has fewer than 25 positive or 25 negative observations. Held-out observations in an unsupported within-domain fold receive no out-of-fold prediction and are omitted from both sides of that judge's ECE and selective-risk contrast. Judges with unsupported outside-domain fits or no retained paired predictions contribute no contrast; ECE averages eligible judges, while risk averages only judges meeting its support rule on both sides, and a cell without an eligible contrast is undefined for that measure.

D.5 Bootstrap and overlap mechanics

H2 and H3 use 10,000 paired BCa draws (Efron, 1987) from one NumPy `PCG64(20260831)` stream consumed in the fixed order H2 macro, H3 ECE macro, H2 domain intervals, and H3 ECE domain intervals. The unsupported ΔRisk measure consumes no draws. Within each draw, artifacts are resampled separately within domain and items are resampled within artifact while keeping complete judge tuples. Undefined domains are omitted; draws with fewer than three valid domains are rejected and replaced from the continuing stream. BCa acceleration is the delete-one-item-within-artifact jackknife with the artifact roster fixed. Bias correction uses strict less-than counting and linear quantile interpolation.

Resampling is independent across artifact–domain cells. Consequently, if one PPE item belongs to several cells, its sampled multiplicity is drawn separately in each cell rather than shared across a membership vector. For H3 only, the fixed IDs of all items in the held-out PPE cell are then excluded from every other PPE training slice, including overlapping memberships, to prevent calibrator leakage. The resulting intervals preserve the implementation's per-cell resampling rule and do not model cross-cell item covariance.

D.6 Estimand and gate diagram

Figure 2 connects same-item aggregation and calibration transport to the registered decision gates, including the distinct PPE overlap rules.

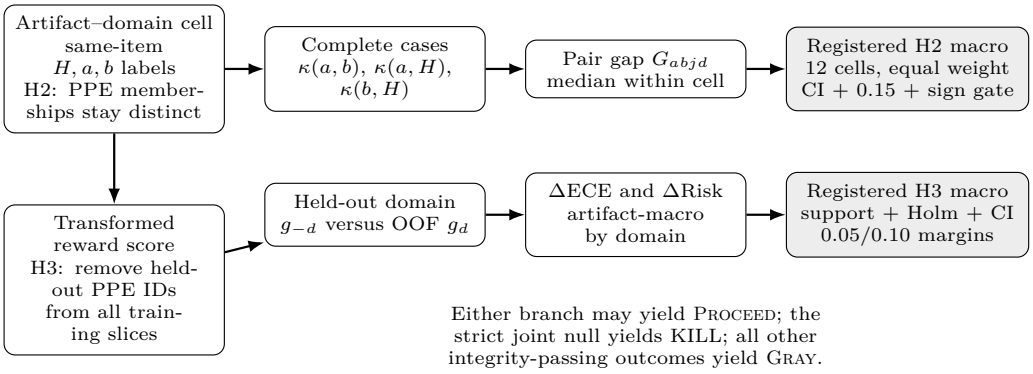

Figure 2: Registered artifact-arm estimands and gates. The diagram separates same-item label aggregation from calibration transport and makes the two PPE overlap rules explicit.

E Sensitivity Analyses and Detailed Diagnostics

E.1 Domain and artifact diagnostic panels

Figure 3 visualizes the H3 estimates in Table 4; Figure 4 shows artifact weights and H2 heterogeneity.

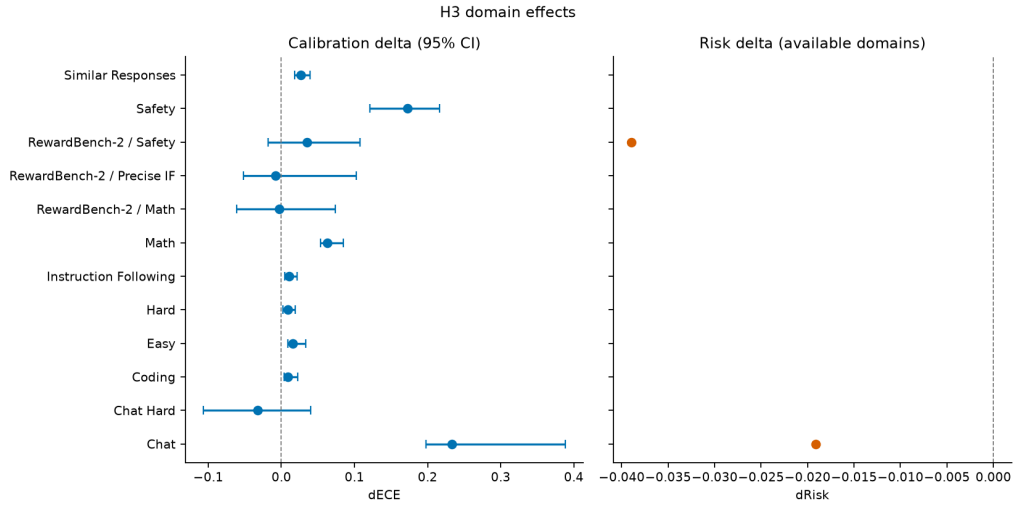

Figure 3: Frozen H3 domain effects. Selective-risk estimates appear only where the registered support rule is met.

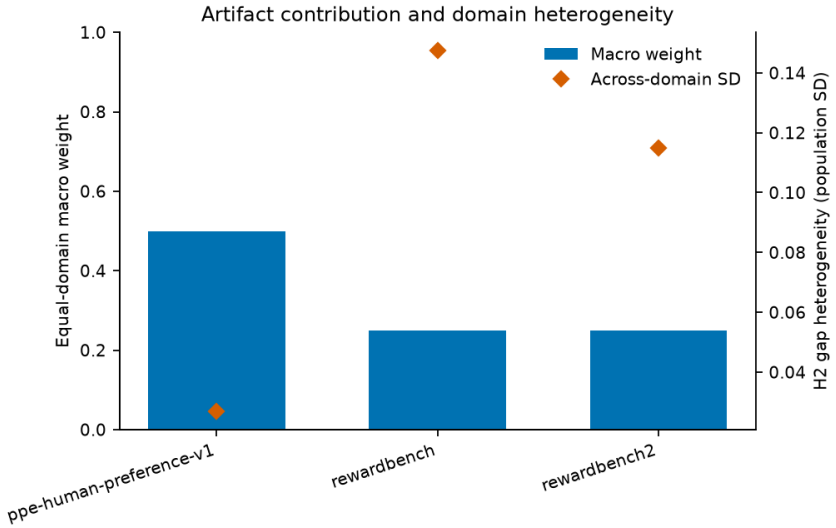

Figure 4: Each artifact's registered macro weight and across-domain H2 heterogeneity.

E.2 Exploratory adapter sensitivity

Section 5.3 reports the exploratory human-tie exclusion analysis and its interpretive limits; Table 6 compares that analysis with the frozen estimand and the additional fixed tie-aware mapping below.

A second sensitivity uses a rule specified in the analysis file header before its first computation: for PPE, emit Tie when the two-way winning softmax score is at most 0.55 (absolute

Table 6: H2 adapter sensitivity alongside the immutable frozen estimand. Only PPE is changed in the exploratory rows.

Estimand / mapping	Macro G
Frozen registered estimand (independent recount)	0.129636
Exploratory: exclude PPE human ties	0.073393
Exploratory: fixed tie-aware PPE judge mapping	0.085889

Table 7: Exploratory reference-noise construction, fixed judge accuracy 0.9 and conditional independence. No empirical reference accuracy is estimated.

Reference accuracy ρ	κ_{JJ}	κ_{JH}	Induced G
0.7	0.64	0.32	0.32
0.8	0.64	0.48	0.16
0.9	0.64	0.64	0.00
1.0	0.64	0.80	−0.16

score difference at most $\log(11/9)$); otherwise retain A/B. Exact ties remain ties and missing scores remain missing. This symmetric neutral band uses no human outcomes or human tie prevalence and is not threshold-tuned. It is a post-review exploratory specification, not part of the original registration. Its score-scale dependence and the fact that Arena "both bad" votes also map to Tie prevent interpreting it as validated human indifference.

E.3 Exploratory reference-noise sensitivity

Consider a balanced latent binary target Y, two symmetric judges with accuracy $a = 0.9$, and a symmetric released reference with accuracy ρ. Assume all three errors are independent conditional on Y. Each observed label then has balanced marginals, so chance agreement is 1/2. The judges agree with probability $a^2 + (1-a)^2$, and each judge agrees with the reference with probability $a\rho + (1-a)(1-\rho)$. Substitution into Cohen's definition (Cohen, 1960) yields the construction

$$\kappa_{JJ} = (2a-1)^2, \quad \kappa_{JH} = (2a-1)(2\rho-1), \quad G = (2a-1)\{(2a-1) - (2\rho-1)\}. \tag{3}$$

This generalizes the external review's counterexample: at $\rho = 0.6$, $\kappa_{JJ} = 0.64$, $\kappa_{JH} = 0.16$, and $G = 0.48$, despite both judges being more accurate than the reference and having independent errors conditional on the target. Table 7 evaluates the requested illustrative reliability grid using Equation 3. These are induced gaps in a hypothetical binary population, not adjusted estimates of the observed multiclass artifact macro. The artifacts do not identify a or ρ relative to an independent latent construct. Construct validity, shared model error, and reference noise cannot be separated by the reported agreement contrast; repeated or independently adjudicated reference evidence would be needed to constrain those explanations.

E.4 Exploratory prevalence-only calibration comparison

The prevalence-only predictor is constant for each fitted calibrator: the weighted fraction of reference-correct labels on exactly the isotonic training split. It uses identical LODO domain weights, PPE overlap exclusions, five hash-defined within-domain folds, and fit-support masks. We evaluate both models on the identical retained judge–item tuples, macro-average judges within each cell, then average the twelve cells. Brier score is $n^{-1}\sum_i (q_i - y_i)^2$; ROC AUC is the probability that a reference-correct item scores above a reference-incorrect item, with half credit for ties. The target throughout is reference agreement, not latent correctness.

The fresh isotonic recount matches every frozen domain ΔECE and the frozen macro 0.044192; prevalence-only macro ΔECE is 0.110376. Uncalibrated transformed-score macro AUC is 0.622939. Isotonic improves Brier score over prevalence only in both evaluations,

Table 8: Exploratory H3 calibration, proper scoring, and discrimination on matched support. Lower ECE/Brier is better; larger AUC is better.

Calibrator	Evaluation	ECE	Brier	AUC
Isotonic	LODO	0.086479	0.228077	0.618226
Isotonic	Within OOF	0.042287	0.214865	0.597249
Prevalence only	LODO	0.117905	0.250065	0.500000
Prevalence only	Within OOF	0.007529	0.222902	0.457932

Table 9: Frozen operating-point summaries at $q \geq 0.80$. Coverage and risk ranges are reported separately across scored judges, including sub-floor points; a risk range omits zero-accepted undefined values. The last column is the registered deterministic top-80%-coverage ΔRisk (LODO minus within-domain), using $\lceil 0.8n \rceil$ items and hash-resolved cutoff ties. RB denotes RewardBench; PPE denotes PPE Human Preference V1; RB2 denotes RewardBench 2.

Artifact–domain	LODO coverage / risk	Within coverage / risk	Top-80% ΔRisk
RB / Chat	0.000–0.318 / 0.009	0.701–0.810 / 0.028–0.117	0.003484
RB / Chat Hard	0.109–0.268 / 0.224–0.500	0.000–0.158 / 0.034	0.000000
PPE / Coding	0.000–0.002 / 0.000–0.429	0.000–0.004 / 0.143–1.000	−0.002395
PPE / Easy	0.000–0.006 / 0.000–0.750	0.000–0.012 / 0.241–1.000	−0.001597
PPE / Hard	0.000–0.004 / 0.000–0.333	0.000–0.003 / 0.083–1.000	−0.002577
PPE / Instruction Following	0.000–0.005 / 0.000–1.000	0.000–0.007 / 0.143–1.000	−0.001408
PPE / Math	0.000–0.007 / 0.000–1.000	0.000–0.005 / 0.071–1.000	−0.003941
RB / Safety	0.366–0.397 / 0.238–0.253	0.131–0.189 / 0.149–0.291	−0.005165
PPE / Similar Responses	0.000–0.002 / 0.000–0.800	0.000–0.002 / 0.071–1.000	−0.000644
RB2 / Math	0.311–0.317 / 0.193–0.241	0.153 / 0.107–0.214	−0.003401
RB2 / Precise IF	0.006–0.019 / 0.333–1.000	0.019–0.056 / 0.222–1.000	0.000000
RB2 / Safety	0.243 / 0.046	0.500 / 0.085	−0.013928
Macro top-80% sensitivity			−0.002631

but within-domain prevalence only has much lower ECE. Thus small ECE alone does not establish useful discrimination. Each LODO constant has AUC 0.5. Within-domain OOF constants differ across folds, so their pooled AUC need not be 0.5; this is an OOF aggregation effect, not within-fold discrimination. The same fold-specific mapping issue affects comparisons of pooled OOF and LODO isotonic AUCs. These diagnostics support the narrower description of H3 as transport of a learned score–correctness mapping, not judges' stated confidence. They do not establish a deployment benefit; selective-use claims remain conditional on supported, comparable-coverage risk estimates.

Fresh code, input checksums, per-pair H2 and per-judge H3 diagnostics, and deliberate-perturbation tests are retained under `pilots/R3/p1/sensitivity/`. Reproduce with `python pilots/R3/p1/sensitivity/analyze.py --check` and `python pilots/R3/p1/sensitivity/test_sensitivity.py`.

E.5 Coverage and operating-point sensitivity

All 12 held-out cells retained 100% of their eligible calibration items after fold-support filtering. Table 9 reports every held-out domain's range across scored judges at the registered $q \geq 0.80$ operating point, including sub-floor points rather than suppressing them. Exact per-judge accepted counts, coverage, risk, and support-validity flags at $q \geq 0.80$ are preserved in `h3.details.*.contrast_item_coverage` of the frozen report. These are operating-point records, not full risk–coverage curves; the protocol's full-curve requirement was not fulfilled by the archived report. The separate top-80% diagnostic supplies one comparable-coverage operating point. A risk range omits only a zero-accepted undefined value; gate validity still requires both at least 25 accepted items and 20% coverage.

The fixed top-80% sensitivity has ample support by construction and remains near zero in every domain; its artifact–domain macro is −0.002631. It is a registered diagnostic,

Table 10: Leave-one-artifact-out reaggregation of frozen H2 domain estimates.

Omitted artifact	Retained domains	Mean H2 gap
PPE human preference	6	0.018122
RewardBench	9	0.148066
RewardBench 2	9	0.185549

Table 11: H2 weighting sensitivity computed from the frozen report. The final two rows are explicitly non-registered and are not passed through the Gray gates. Artifact weights are ordered RewardBench / PPE / RewardBench 2.

Summary	Effective artifact weights	H2 gap
Registered cell-equal macro	25.00% / 50.00% / 25.00%	0.129636
Non-registered: artifact equal	33.33% / 33.33% / 33.33%	0.092465
Non-registered: item pooled	6.37% / 89.22% / 4.41%	0.204148

not a replacement for the $q \geq 0.80$ gate estimand, whose two-valid-domain status remains `GRAY-null`.

E.6 Leave-one-artifact-out diagnostic

Table 10 gives the deterministic artifact-deletion diagnostics discussed in Section 5.3; these reaggregations have no separate bootstrap interval or gate.

E.7 Non-registered weighting sensitivity

Table 11 reaggregates the frozen report without changing any cell estimate. These are deterministic, post-registration sensitivities with no bootstrap interval and no gate status. "Artifact equal" first averages cells within each artifact and then assigns each artifact one-third weight. "Item pooled" first pools each artifact's stored confusion matrices across its domains (using the report's reporting-only PPE Overall matrix), computes one artifact-wide H2 gap, and then weights those gaps by the 1,145, 16,038, and 793 unique artifact items. The latter avoids counting overlapping PPE domain memberships more than once, but it gives PPE 89.22% of the summary. Thus the registered macro lies between two substantively different target populations; neither sensitivity replaces its cell-equal estimand.

E.8 Worked disagreement matrix

One frozen PPE Coding matrix makes the gap concrete. Across 2,954 complete items, Athene-RM-70B and Athene-RM-8B agree on $1{,}329 + 1{,}253 + 2 = 2{,}584$ labels (87.47%; $\kappa = 0.750178$). Against the human vote, their respective diagonal counts are only $627 + 594 + 4 = 1{,}225$ (41.47%; $\kappa = 0.152875$) and $603 + 585 + 2 = 1{,}190$ (40.28%; $\kappa = 0.135619$). The resulting pair gap is 0.605931, even though the Coding cell median across all 91 pairs is 0.260815. This is a matrix-level worked case, not a reconstructed individual item: the frozen report stores the three pairwise confusion matrices but not a three-way item identifier in that summary record.