When a Reliability Coefficient Orders References Backwards: Identifiability, a Calibrated Test, and Design Rules for the Human References of LLM Judges

Anonymous authors
Paper under double-blind review

Abstract

LLM judges are validated by checking that they agree with a panel of human raters, and the panel is trusted when its reliability coefficient is high. We show that this can go wrong in two ways. First, a judge can pass validation by fitting one panel. When two independent panels rate the same items, their covariance measures what they agree on: the *common signal*, which is not necessarily ground truth. Under an additive model, and for a judge whose errors are independent of the reference's, this sets a validity ceiling on how well any judge can agree with a panel through that signal. On SummEval, G-Eval's published correlations with the expert panel (0.455–0.582) exceed the upper confidence bound of this ceiling (at most 0.30). Recomputed from its released outputs, G-Eval correlates 0.50–0.57 with the experts and only 0.02–0.06 with the crowd. Nine recent judges show the same split in 33 of 36 judge–dimension pairs, one of them borderline. Second, the reliability coefficient can rank references backwards. A component shared only within one panel raises its reliability and lowers its fidelity at the same time. In 19 of 28 D3code region pairs, the panel that looks more reliable is the less faithful one. We give a per-panel test for this component, with 4.8% false positives at a nominal 5% and 93.4% power at $\tau^2 = 0.3$. Without additional information, no one- or two-way variance-components model of a single panel can separate the component from signal, and the designs that maximise reliability or fidelity cannot be checked. A checkable design costs only $O(1/b)$ in fidelity. Only five of the fourteen released references we survey allow the check. Identifying the common signal from two panels is classical, and normalized cross-replication reliability already reads it as a disattenuated agreement. What is new is the per-panel test and what it reveals about how LLM judges are validated.

1 Introduction

On SummEval coherence, G-Eval agrees with the expert panel at 0.57 and with the crowd panel at 0.02, on the same summaries. Both panels report high reliability ($\rho_k = 0.79$ and 0.83). So what did the validation check?

LLM-as-a-judge studies use human ratings as the reference: a judge is trusted if it agrees with people. The quality of that reference is usually summed up by one reliability number. That number depends on how ratings are aggregated, and it does not show whether the reference measures what it is meant to measure. We start with what it does to a decision. It can put two references in the wrong order, and a judge can agree with one of them far beyond anything a second panel confirms. The designs that maximise the number (one large panel) or the reference's fidelity (many single-rater panels) are exactly the ones that cannot be checked (§3.3). The number is *not self-explanatory*: the same value supports opposite conclusions, depending on facts that are almost never reported.

The first problem is the *unit*. Agreement statistics measure the reliability of a *single* rating, ρ_1. A benchmark uses the *mean* of k ratings, whose reliability ρ_k is higher and grows with k (Spearman–Brown). The two share a name, and they can rank panels, and the judges tested against them, in opposite orders. That can change which system, or which LLM judge, a study declares better.

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.

The second problem is *validity*. Even ρ_k mixes two things: the signal the panel should track, measured by its fidelity $\rho_p^\star = \sigma_s^2/\mathrm{Var}(\bar{y}_p)$, and a component τ_p^2 shared by the panel's raters that averaging cannot remove. We call τ_p^2 the panel-specific component. It is defined *relative to the common signal*, the part independent panels agree on, which need not be true quality. It can be method error, a difference in construct or population between panels, or both. Our test therefore certifies reliability *relative to that common signal*; it does not claim to recover true quality.

Contributions. We report three findings. (i) *A judge can pass validation by fitting one panel.* Fidelity gives each panel a validity ceiling $\sqrt{\rho^\star}$, distinct from the noise ceiling $\sqrt{\rho_k}$. On SummEval, an LLM judge exceeds the *upper confidence bound* of this ceiling. Like nine recent judges in 33 of 36 judge–dimension pairs, it tracks the expert panel but not the crowd; we call this *panel fitting* (§4). (ii) *The reliability coefficient can rank references backwards.* A panel-specific component raises ρ_k and lowers $\rho^\star$. The panel that looks more reliable is the less faithful one in 19 of 28 D3code region pairs. On Jigsaw, the point estimates rank the three pools in reverse order, with two of the three pairs significant (§4, Appendix J). (iii) *The designs that maximise reliability or fidelity cannot be checked.* One panel cannot separate its shared component from signal (Prop. 1), and single-rater panels leave nothing to test (Prop. 3). A checkable design costs only $O(1/b)$ in fidelity (§3.3). Behind these findings are four tools. The first is the validity ceiling. The second is a per-panel test with a stated false-positive rate (§3.2). The third is a reporting protocol (§3.4), and the fourth is a set of six algorithms with an audit card (Appendix D). Only five of the fourteen released references we survey allow the test (Appendix M), so we offer the two-panel design as a rule for new references.

The first step, identifying the common signal from two panels, is classical (§2). What is new is what it shows about how LLM judges are validated.

2 Related work

Reliability of aggregated human annotations. Reliability depends on the unit being measured. Wong & Paritosh (2022) stress that the reliability of one rating and of an aggregate are different quantities, and that an evaluation should report the level it consumes. Classical work distinguishes several intraclass correlation coefficients and their assumptions (Shrout & Fleiss, 1979; Koo & Li, 2016). We make this explicit through ρ_1 and ρ_k. We then ask a separate question: whether the aggregate is identifiable as a measure of what independent panels share.

Variance components and construct validity. Generalizability theory splits rating variance across facets (Cronbach et al., 1972), and the multitrait–multimethod framework shows that agreement can come from shared method variance (Campbell & Fiske, 1959). Our random-effects model is a restricted instance of the former. What we need is the difference between a component shared by *all* panels and one shared only *within* one. That difference turns $\rho^\star$ into a question of identifiability (Props. 1–2). A component common to one panel's raters need not be noise, but it cannot be equated with the target signal either. Our decomposition identifies a statistical component under stated assumptions. It does not recover ground truth.

Cross-replication reliability. The closest precedent is cross-replication reliability (xRR; Wong et al., 2021). It also places several rater pools on the *same* items and reports a chance-corrected cross-pool *agreement* coefficient κ_x. Wong et al. go further than raw agreement. They define a *normalized* $\kappa_x/\sqrt{\mathrm{IRR}_X\,\mathrm{IRR}_Y}$, explicitly modelled on Spearman's correction for attenuation (Spearman, 1904), and read it as a disattenuated correlation between two pools. They also document the phenomenon we formalise: pools that agree well internally but barely with each other (for *contentment*, internal reliabilities of 0.45 and 0.64 against a cross-pool κ_x of -0.03). So we claim neither the identification nor the phenomenon. We add three things. The quantity is decomposed *per panel*, not reported as a symmetric pair statistic, and at the finite k a benchmark actually uses. The panel-specific component comes with a test whose size and power we report; Wong et al. list confidence intervals and hypothesis tests as future work. And the resulting per-panel ceiling is applied to a machine evaluator, which that line of work does not do. Calibration matters in practice. For two expert pools recruited by Dutta et al. (2023), normalized xRR reaches 1.073. A correlation cannot exceed 1. Without a calibrated test, a reader cannot tell whether this excess is noise.

Dependence among LLM judges. Another line studies dependence among automatic evaluators. Kim et al. (2025) document correlated errors across many models; Yang et al. (2026) treat evaluator replacement as a measurement-validity problem and find that juries add little when errors are correlated; and Kohli (2026) finds that a nine-judge panel carries only a few effective votes. These works correct dependence *among automatic judges* to recover judge *accuracy*. Concurrent work by Sunkavalli (2026) studies an anchor–judge model. It gives a closed-form estimator and a set of diagnostics, with 5% thresholds calibrated under the null and item-bootstrap intervals. Like us, it notes that identifiability is not new relative to MTMM method-factor models. Its shared component sits on the *judge* side. Ours is a property of the human panels that judges are validated against, and it yields a ceiling on judge–human correlation that an anchor–judge model does not define. Calderon et al. (2025) give a leave-one-annotator-out test with a Benjamini–Yekutieli false-discovery guarantee. It works *within* one pool and treats that pool as the target, so a judge that fits the pool's shared idiosyncrasy passes. We ask the prior question: can the human reference itself be audited?

Position of this work. Identifying a common signal from two measures is old (Spearman, 1904; Campbell & Fiske, 1959), and xRR already reads cross-pool agreement as a disattenuated correlation. We use that step and do not claim it (Table 12, Appendix L). What we add is what it shows for the human references of LLM judges: a ceiling that a published judge exceeds, a reliability coefficient that ranks references backwards, and a design rule with its price. The classical literature asks whether two human panels agree. We ask whether a machine evaluator, validated against one panel, tracks anything a second panel would confirm. On SummEval it does not.

3 METHOD

We validate each proposition numerically as it is introduced; the dataset audits follow in §4.

Notation. p indexes panels, i items and j raters. The common item-level signal is μ_i, with variance σ_s^2. Each panel p adds a panel-specific shared component $c_{p,i}$, with variance τ_p^2. Each rating adds idiosyncratic noise $\varepsilon_{p,i,j}$, with variance σ_p^2. We write $\bar{y}_{p,i}$ for panel p's mean rating of item i. ρ_1 is the reliability of one rating and ρ_k that of the mean of k ratings. Panel p's fidelity to the signal is $\rho_p^\star = \sigma_s^2/(\sigma_s^2 + \tau_p^2 + \sigma_p^2/k_p)$. P, k and n are the numbers of panels, raters per panel and items. The budget per item is $b = Pk$, and $\lambda = \tau^2/\sigma^2$ is the ratio of shared to idiosyncratic variance.

3.1 PRELIMINARIES AND CLASSICAL RELIABILITY

Let a panel of k raters score n items. Under a one-way model $y_{ij} = \mu_i + \varepsilon_{ij}$ with $\mathrm{Var}(\mu) = \sigma_s^2$ and $\mathrm{Var}(\varepsilon) = \sigma_e^2$,

$$\rho_1 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}, \qquad \rho_k = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2/k} = \frac{k\,\rho_1}{1 + (k-1)\rho_1},$$

the second being Spearman–Brown. A benchmark that scores by the panel *mean* $\bar{y}_i$ has reliability ρ_k, not ρ_1. When a judge's error is independent of the reference error, the *Pearson* correlation between the judge and $\bar{y}_i$ obeys the attenuation bound $|\,\mathrm{Corr}(J, \bar{y}_i)| \le \sqrt{\rho_k}$. This is a variance-based noise ceiling, and it rises with k. Our correlations in §4 are Spearman correlations. We use $\sqrt{\rho_k}$ as their ceiling too, assuming the judge–reference relationship is monotone enough for Pearson and rank ceilings to be comparable. The bound is exact for Pearson and an approximation for Spearman. Reporting ρ_1 while consuming $\bar{y}_i$ reports a different quantity from the one the benchmark uses, and comparing panels (or judges on different panels) on different units invites reversals.

3.2 IDENTIFIABILITY AND A CALIBRATED DEPENDENCE TEST

Model. We model P panels rating the same n items:

$$y_{p,i,j} = \mu_i + c_{p,i} + \varepsilon_{p,i,j},$$

The rating has three parts. The signal μ_i has variance σ_s^2. The panel-specific *shared component* $c_{p,i}$ has variance τ_p^2; it is i.i.d. across items and independent across panels. The idiosyncratic noise $\varepsilon_{p,i,j}$ has variance σ_p^2. The panel mean $\bar{y}_{p,i} = \mu_i + c_{p,i} + \bar{\varepsilon}_{p,i}$ has error variance $\tau_p^2 + \sigma_p^2/k_p$. The shared

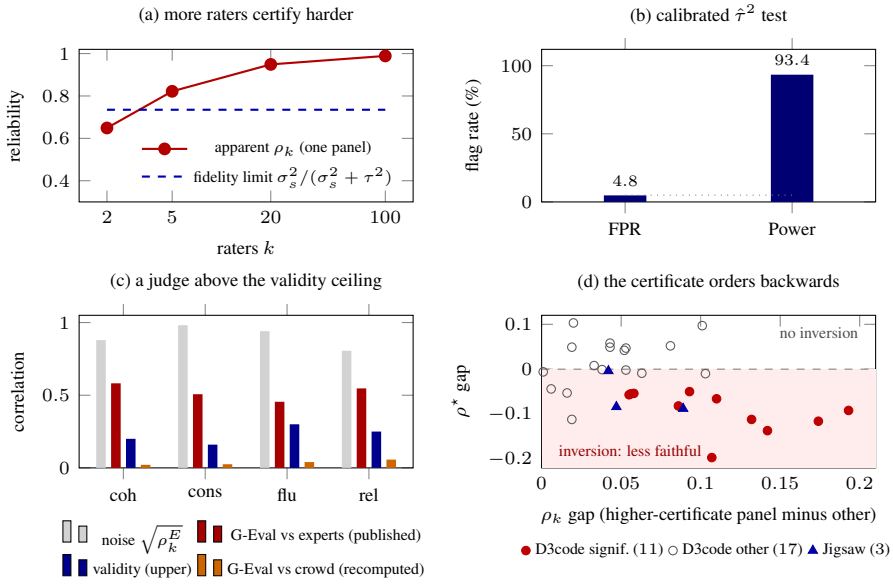

Figure 1: (a), (b) **[Synthetic]**; (c), (d) **[Real data]**. (a) Prop. 1: $\hat{\rho}_k$ inflates toward 1 with more raters while fidelity stays capped (analytic, $\sigma_s^2 = 1$, $\tau^2 = 0.36$, $\sigma^2 = 1.47$). (b) The $\hat{\tau}^2$ test: 4.8% false positives at a nominal 5%, 93.4% power at $\tau^2 = 0.3$. (c) SummEval: G-Eval's published correlations with the expert reference sit below the noise ceiling but *above* the upper confidence bound of the validity ceiling in every dimension, while its recomputed correlation with the crowd panel is near zero: panel fitting. (d) Each point is a panel pair, oriented so that the first panel has the higher certificate ρ_k; below zero, that panel has the *lower* fidelity $\rho^\star$. This holds for 19 of 28 D3code region pairs (11 significant) and all three Jigsaw pairs (two significantly). Gaps use each pair's own covariance (Appendix J).

term does not shrink as raters are added. The common signal μ_i is defined operationally as what independent panels share, not as ground-truth quality, and all our claims concern the fidelity $\rho_p^\star$ to it.

Proposition 1 (One panel is not enough to identify the shared component)**.** *From a single panel's ratings, σ_s^2 and τ_p^2 enter every moment only through the between-item variance $\sigma_s^2 + \tau_p^2$ and the within-item variance σ_p^2; hence they are not separately identifiable, and the fidelity $\rho_p^\star$ is not identifiable. A two-way (item×rater) model does not help: $c_{p,i}$ depends on the item alone, so it aliases with the item main effect, not with any rater or interaction facet. The apparent aggregate reliability $\hat{\rho}_k = (\sigma_s^2 + \tau_p^2)/(\sigma_s^2 + \tau_p^2 + \sigma_p^2/k) \to 1$ as $k \to \infty$, while $\rho_p^\star \to \sigma_s^2/(\sigma_s^2 + \tau_p^2) < 1$.*

So collecting *more* raters makes a benchmark's reliability certificate more confident while widening its gap to fidelity (Fig. 1a). No estimator that uses only one panel's rating matrix can separate these components under this model class without extra identifying information. A better model does not help. External ground truth or strong parametric assumptions could identify them; a single panel cannot.

Proposition 2 (Two independent panels identify the signal variance)**.** *If $P \geq 2$ panels have mutually independent shared errors, then for any $p \neq q$, $\mathrm{Cov}(\bar{y}_p, \bar{y}_q) = \sigma_s^2$, so the* common signal variance *σ_s^2 is identified, and hence each fidelity $\rho_p^\star = \sigma_s^2/\mathrm{Var}(\bar{y}_p)$. The panel-specific shared-error variance τ_p^2 is identified only when the idiosyncratic variance σ_p^2 is estimable, e.g. from two raters per panel ($k \geq 2$; Prop. 3). For $P > 2$ the $\binom{P}{2}$ pairwise covariances all estimate σ_s^2, overdetermining it and adding robustness to a contaminated panel.*

In simulation, $\mathrm{Cov}(\bar{y}_1, \bar{y}_2)$ recovers σ_s^2 to within 0.4% for every τ. Over the same range, $\hat{\rho}_k$ climbs from 0.65 to 0.99, while fidelity stays below its limit of 0.735.

A calibrated dependence test. Proposition 2 turns detection into estimation, in two steps. First, take the between-item variance of the panel means minus the cross-panel covariance:

$$\hat{g}_p = \widehat{\mathrm{Var}}(\bar{y}_p) - \widehat{\mathrm{Cov}}(\bar{y}_p, \bar{y}_q), \qquad \mathbb{E}[\hat{g}_p] = \tau_p^2 + \sigma_p^2/k_p.$$

This is the panel mean's total *non-common* error: shared error plus residual noise. Second, subtract the within-panel idiosyncratic term, which can be estimated when $k \geq 2$. What remains is the shared component,

$$\hat{\tau}_p^2 = \hat{g}_p - \hat{\sigma}_p^2/k_p, \qquad \mathbb{E}[\hat{\tau}_p^2] = \tau_p^2,$$

which is zero when panel p has no panel-specific component ($\tau_p^2 = 0$) and positive otherwise. This null, $H_0 : \tau_p^2 = 0$, is about shared error *within* a panel. It is distinct from the independence assumption used to identify σ_s^2, which is about shared error *between* panels. When σ_p^2 cannot be estimated credibly, we report $\hat{g}_p$ as "residual panel error" rather than as shared-error variance. The test resamples *items* with replacement (B bootstrap draws, recomputing everything each time). It forms the percentile interval of $\hat{\tau}_p^2$ and flags the panel when the lower bound exceeds zero, a one-sided test of H_0 against $H_1 : \tau_p^2 > 0$. The *unconstrained* estimator can be negative in finite samples, through sampling variation or model departure. We calibrate the test by simulation. In our simulation the false-positive rate is 4.8% at a nominal 5%, and the power against $\tau^2 = 0.3$ is 93.4% (Fig. 1b; simulation parameters in Appendix K). With ratings rounded to a 1–5 scale, with unequal numbers of raters, or with both, the false-positive rate stays near the nominal 5% (at most 6.8% in 400 replications, whose Monte Carlo standard error is about 1.3 points). Unlike a qualitative checklist, the test states how often it is wrong (Algorithm 2).

Consequences. With two panels the test *detects* dependence but does not *resolve* it. A flagged reference is reported as flagged, because the shared component cannot be identified within either panel (Prop. 1). This gives a design rule. A benchmark that must support dependence-robust reliability claims needs at least two independent panels on shared items. Adding raters to one panel is no substitute, and the number of items sets the detection power (Appendix F; Algorithm 5).

3.3 Reference construction under a fixed budget

Two goals, two allocations. Fix a budget of $b = Pk$ ratings per item. Fidelity, $R(P,k) = \sigma_s^2/(\sigma_s^2 + \tau^2/P + \sigma^2/b)$, increases strictly in P. It is maximised at $P = b$, $k = 1$: all the budget goes into independent single-rater panels (at $b = 24$, R rises from 0.8392 at $P = 2$ to 0.9486 at $P = b$). This R is the fidelity of the *grand mean over all P panels*, where the shared component averages down as τ^2/P. A benchmark that uses a single panel's mean keeps the full τ^2 and never reaches this value, so the tension concerns the audit design, whose panels then serve as the reference. Detection follows a different law. For the pooled estimator of a common τ^2, $\mathrm{Var}(\hat{\tau}^2) = m(\tau^2 + \sigma^2/k)^2/(n(P-1))$, with $m = 2$ under Gaussianity (Prop. 5, Appendix F). It needs within-panel replication, so its optimum lies away from the fidelity optimum (Fig. 2). For small λ the optimum is interior; for large λ it sits at $k = 2$. It is never at $k = 1$. Our test uses the per-panel estimator, whose variance is larger. At $b = 24$, its detection SNR is nearly flat between two and four panels, whatever λ, and its optimum is never at $k = 1$ either (Appendix F). This analysis guides the design of a reference; the audit itself flags panels one at a time (Algorithm 2).

Proposition 3 (Detectability excludes the fidelity optimum, at a cost of $O(1/b)$)**.** *Under the proposed random-effects decomposition, and without external information about the individual-rater variance σ^2, a single rating per panel ($k = 1$) does not identify the shared-error component: with $k = 1$ the terms $c_{p,i}$ and $\varepsilon_{p,i,1}$ are indistinguishable, so the design provides no within-panel replication to separate the two error levels, for any P; at least two raters per panel are required. So the fidelity-optimal allocation ($P = b$, $k = 1$) falls outside the detectable region, which is $k \geq 2$, i.e. $P \leq b/2$. The fidelity cost of the best auditable design is $R(b,1) - R(b/2,2) = \tau^2/(\sigma_s^2 b) + O(1/b^2)$.*

This is a property of *this* design. It is not a general impossibility. External replicate measurements or strong parametric assumptions could identify σ^2 at $k = 1$. Still, the allocation that maximises the reliability coefficient removes exactly the replication needed to check it. A trustworthy protocol therefore keeps $k \geq 2$ raters in at least two independent panels. For per-panel flags, two to four panels give useful power in the simulated settings (Algorithm 4); a pooled estimate of a common τ^2 is best served by P near $\lambda b + 2$.

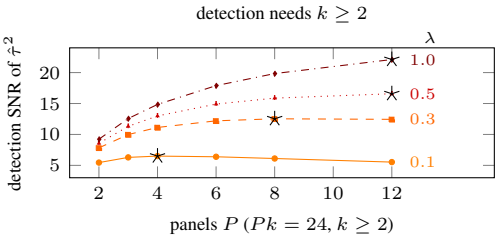

Figure 2: **[Synthetic]** Detection SNR of the pooled $\hat{\tau}^2$ (Prop. 5) at budget $Pk = 24$ peaks at $k \geq 2$ (stars), never at the fidelity optimum $k = 1$ (Prop. 3); entering the detectable region costs 1.1 points of fidelity (0.9486 vs. 0.9375). Parameters in Appendix K.

The price of auditability. The tension can be measured. At a fixed budget b, the gap between the fidelity optimum and the best auditable design is

$$R(b,1) - R(b/2,2) = \frac{\sigma_s^2\,\tau^2/b}{D_1 D_2}, \qquad D_1 = \sigma_s^2 + \frac{\tau^2+\sigma^2}{b}, \quad D_2 = D_1 + \frac{\tau^2}{b},$$

which is $\tau^2/(\sigma_s^2 b)$ to leading order. The price of auditability falls as $O(1/b)$. With the parameters of Fig. 2 ($\sigma_s^2 = 1$, $\tau^2 = 0.3$, $\sigma^2 = 1$, $b = 24$) it is 1.1 percentage points of fidelity (0.9486 vs. 0.9375).

3.4 The reporting protocol

Inverting Spearman–Brown gives the smallest panel, $k_{\min}$, that reaches a target ρ_k^{tgt} (Appendix F). We propose three declarations for any human reference: (1) *which quantity* is reported (ρ_1 or ρ_k); (2) *how many ratings* k the benchmark consumes, with $k_{\min}$ for the target; (3) *which scale or unit*, and the object spread. Figures that are compared must use the same unit. Where two panels exist, the dependence test should accompany any ranking claim. Algorithm 1 (Appendix D) packages these steps; Algorithms 2–6 specify them as functions, with a practical guide. Unlike current threshold conventions, Algorithm 1 comes with *stated operating characteristics*: whether the coefficient can be checked under the design, and with what false-positive rate.

How often is the test applicable? Among human-reference datasets that release *rater-level* votes (Appendix M), only five provide two independent panels on the same items as released: SummEval, DICES-990, D3code, MultiPref and IRep. The Jigsaw pools are a sixth, stress-test case, under an independence caveat. Most others give a single multi-annotated pool or a single annotation per item. The auditable design is the exception, which is the argument for building it into reference collections.

4 Experiments

Datasets. Five released datasets allow the test as released (Appendix M). Four of them come from LLM and NLP evaluation, and we analyse them in full. SummEval has three experts and five crowd workers on 1,600 summaries. DICES-990 (Aroyo et al., 2023) has a US pool and an India pool, with about 36 ratings per item from each, on 990 conversations. D3code (Mostafazadeh Davani et al., 2024) has eight geo-cultural regions on 4,554 items. MultiPref (Miranda et al., 2025) has two non-expert and two expert annotators per item, on 10,461 comparisons; the two groups share no annotator. IRep ($k = 2$) and Jigsaw (shared raters) serve as stress tests. Source files and preprocessing for each are in Appendix K.

The certificate cannot tell auditable from unauditable. Panels with similar or high certificates can share almost none of their item-level signal, or almost all of it. On three of our two-panel datasets the cross-panel correlation of item means is $r \approx 0.04$ (SummEval coherence), $r \approx 0.3$ (D3code regions) and $r \approx 0.83$ (DICES US/India). The dependence test turns this into a panel-specific $\hat{\tau}^2$ and a fidelity $\rho^\star$ (Table 1; MultiPref, whose two-rater panels are only modestly reliable, comes below). Two of these three contrasts (DICES, D3code) compare *populations* rather than

Table 1: **[Real data]** Reported reliability does not predict the audit outcome: panels with similar or high certificates can share as little as 3% or as much as 93% of their reliable variance. Cross-panel correlation r of item means, the flagged panel-specific component $\hat{\tau}^2$, fidelity to the common signal $\rho^\star$, and the shared fraction $\sigma_s^2/(\sigma_s^2 + \hat{\tau}^2)$, on three real two-panel datasets native to LLM/NLP evaluation. ρ_k, $\hat{\tau}^2$ and $\rho^\star$ are those of the flagged panel (SummEval: experts; D3code: North America; DICES: US). $\hat{\tau}^2$ is on each dataset's raw rating scale and is not comparable across rows.

dataset	contrast	r	ρ_k	$\hat{\tau}^2$	$\rho^\star$	shared
SummEval (coherence)	expert vs. crowd	0.04	0.79	0.82	0.027	3%
D3code (N.Am. vs. Sinosphere)	culture	0.29	0.50	0.24	0.24	48%
DICES-990 (US vs. India)	population	0.83	0.90	0.002	0.84	93%

methods. A flagged $\hat{\tau}^2$ there is a cultural or population difference. It is legitimate, but it changes how the reference should be read, and the single coefficient hides it.

Certificate–fidelity inversion: the coefficient orders references backwards. Two panels on the same items share one covariance c, so $\rho_p^\star = c/\mathrm{Var}(\bar{y}_p)$ and $\rho_p^\star/\rho_q^\star = \mathrm{Var}(\bar{y}_q)/\mathrm{Var}(\bar{y}_p)$. A panel-specific component τ_p^2 enlarges $\mathrm{Var}(\bar{y}_p)$. That *raises* ρ_k and *lowers* $\rho^\star$. The certificate prefers the panel that is less faithful to what the two panels share. In 19 of the 28 D3code region pairs, the panel with the higher ρ_k has the lower $\rho^\star$. In 11 of them, both differences exclude zero with per-pair intervals, and 8 remain after a Bonferroni correction over the 28 pairs. The ratio is an identity; what the data show is how often the two orders disagree. On Jigsaw, the order by ρ_k (LGBTQ 0.646, Control 0.604, African American 0.557) is the exact reverse of the order by $\rho^\star$ (African American 0.598, Control 0.551, LGBTQ 0.548). Two of the three pairs are significant; the Control–LGBTQ fidelity gap is 0.003. The pool ranked last by its certificate is ranked first by fidelity (Fig. 1d; Appendix J). A pair can also reverse when the reported unit changes. On SummEval coherence the experts lead on ρ_1 (0.559 vs. 0.494), and the crowd leads on the ρ_k the benchmark consumes (0.830 vs. 0.792).

Noise ceilings. We take each metric's summary-level Spearman correlation with the three-expert reference from Table 1 of Liu et al. (2023). They follow the protocol of Zhong et al. (2022), whose human scores are the expert annotations. Under that protocol, a summary-level correlation is the average of per-document correlations over the 100 source documents. We therefore compute every ceiling on document-centred panel means, with a document bootstrap. The ceiling for the three-expert reference is $\sqrt{\rho_k^E}$. It is a Pearson ceiling, used as an approximation for Spearman (§3.1): 0.879 (coherence), 0.982 (consistency), 0.941 (fluency) and 0.806 (relevance). The attainable correlation thus differs by dimension and, since ρ_k grows with k, by panel size. Relative to the ceiling (Appendix C), the overlap and embedding metrics (ROUGE, BERTScore, MoverScore) reach only 11–40%. G-Eval reaches 48–68%, highest or tied-highest: closest on relevance (68%) and coherence (66%), furthest on fluency (48%).

Reliability ceiling vs. validity ceiling. These are *reliability* ceilings. $\sqrt{\rho_k^E}$ is the largest correlation any predictor can have with the noisy mean of k experts. The *validity* ceiling $\sqrt{\rho^{\star E}}$ is the largest Pearson correlation a judge can have with the expert panel mean through the *common signal*, that is, through what an independent panel also measures. It holds under the additive model, for a judge whose errors are independent of the reference's. Computed the same way on document-centred means, it is far lower on SummEval: 0.06 (coherence), 0.00 (consistency, point estimate below zero), 0.22 (fluency) and 0.15 (relevance). The intervals are wide because σ_s^2 is small; their upper bounds are 0.20, 0.16, 0.30 and 0.25. Three of the four intervals include zero, so we compare against the *upper* bound, the conservative choice. G-Eval's published summary-level correlations with the expert mean (Liu et al., 2023) are 0.582, 0.507, 0.455 and 0.547. They *exceed the upper bound of the validity ceiling in every dimension* (Fig. 1c). The attenuation bound is exact for Pearson and approximate for Spearman, but the gap (0.455–0.582 against at most 0.30) is far too large for that approximation to close. With Pearson on pooled summaries, where the bound is exact, G-Eval reaches 0.58–0.61 with the experts (Appendix C). For a judge whose errors are independent of the

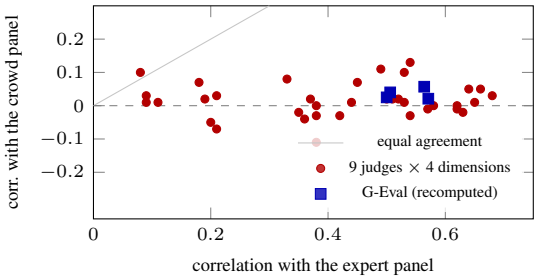

Figure 3: **[Real data]** SummEval: 39 of the 40 points lie below the diagonal (the exception is Mistral Small on fluency, 0.08 vs. 0.10), and the expert–crowd gap excludes zero in 33 of the 36 judge–dimension pairs. Agreement with the expert panel reaches 0.68, while agreement with the crowd panel stays within $[-0.11, 0.13]$. Red: nine judges (Guerdan et al., 2025), four dimensions each, item-level Pearson on 200 summaries. Blue: G-Eval recomputed, summary-level Spearman on all 1,600 (Appendix I).

expert-specific component, the part of its correlation that runs through the common signal cannot exceed $\sqrt{\rho^{\star E}}$. The excess is agreement with structure that no independent panel confirms. This is *panel fitting*. We use the term as a descriptive label for agreement that tracks one panel beyond what the common signal supports. It says nothing about the judge's intent or training. Normalising a judge by the usual noise ceiling $\sqrt{\rho_k}$ certifies agreement with a *panel*; $\sqrt{\rho^\star}$ shows how little of that agreement an independent panel shares. The decomposition cannot tell whether the expert panel is idiosyncratic or the crowd measures a different construct; the crowd's own ceiling (0.08, 0.00, 0.22, 0.17) is comparable. The judge's agreement is not confirmed by the only independent panel available, and that confirmation is what a validation certificate is taken to provide.

A direct test: judges against both panels. Recomputed from its released responses, G-Eval's summary-level correlation is 0.50–0.57 with the expert panel and 0.02–0.06 with the crowd panel (Figs. 1c and 3; Appendix I). Nine recent judges (Guerdan et al., 2025) show the same split on 200 summaries: up to 0.68 with the experts, and between -0.11 and 0.13 with the crowd. The per-pair interval of the gap excludes zero in 33 of the 36 judge–dimension pairs. On MultiPref, whose panels share almost all their reliable variance, GPT-4 agrees with both ($r = 0.33$ and 0.37). A low ceiling is a fact about the pair of reference panels, not about the judge. To validate a judge, compare it with *both* panels (Algorithm 3).

SummEval. We apply the test to SummEval (Fabbri et al., 2021), where two disjoint panels (three experts and five crowd workers) rate the same 1,600 summaries. Our one-way recomputation gives, for coherence, $\rho_1 = 0.559$ (experts) and 0.494 (crowd), and $\rho_k = 0.792$ and 0.830 (Table 2). The panel-specific component $\hat{\tau}^2$ is positive for both panels in every dimension (0.33 to 0.82, all 90% intervals excluding zero), so ρ_k exceeds cross-panel fidelity. The two panels' item means are almost uncorrelated ($r \leq 0.07$ in every dimension), and fidelity to the common signal nearly vanishes ($\rho^\star = 0.005$–0.085; Table 2). A certificate of ρ_k up to 0.96 sits on a shared signal of a few percent. For consistency it is about 0.5%. Haldar & Hockenmaier (2025) find the same: expert–crowd Krippendorff's α on SummEval is at most 0.247, well below either panel's internal reliability.

DICES, D3code and MultiPref. On DICES-990 (Aroyo et al., 2023) the US and India pools agree well ($r = 0.83$, 93% shared), and each aggregate tracks the common signal ($\rho^\star \approx 0.84$ and 0.81). The test still flags a small population component in both pools ($\hat{\tau}^2 \approx 0.002$). Here the reference is both reliable *and* auditable. On D3code (Mostafazadeh Davani et al., 2024) the regional pools agree only moderately ($r = 0.26$–0.47 over 28 pairs), and the test discriminates. For North America vs. Sinosphere it flags North America ($\hat{\tau}^2 = 0.24$, $\rho^\star = 0.24$, 48% shared) but not Sinosphere, and at least one pool is flagged in every pair. MultiPref (Miranda et al., 2025) is preference data. We checked that its non-expert and expert pools are disjoint (89 and 138 identifiers, none shared). Its certificate is low ($\rho_k \approx 0.42$–0.45 at $k = 2$) but almost entirely shared signal, and the test flags a small non-expert component on overall preference ($\hat{\tau}^2 = 0.052$). Full results are in Appendix H.

Table 2: **[Real data]** SummEval per dimension. Left: one-way ρ_1/ρ_k per panel and fidelity to the common signal $\rho^\star$ (note ρ_k up to 0.96 over $\rho^\star$ between 0.005 and 0.085). Right: $\hat{\tau}^2$ with 90% item-bootstrap interval (all exclude zero). The $\rho^\star$ intervals are wide because σ_s^2 is small (Appendix C); the validity ceiling, recomputed on document-centred means, is analysed below.

	reliability & fidelity to common signal						$\hat{\tau}^2$ (90% CI)	
dimension	ρ_1^E	ρ_1^C	ρ_k^E	ρ_k^C	$\rho^{\star E}$	$\rho^{\star C}$	expert	crowd
coherence	0.559	0.494	0.792	0.830	0.027	0.058	0.82 [0.77,0.88]	0.39 [0.34,0.45]
consistency	0.899	0.495	0.964	0.830	0.005	0.007	0.81 [0.71,0.91]	0.45 [0.40,0.49]
fluency	0.726	0.483	0.888	0.824	0.050	0.053	0.45 [0.38,0.51]	0.39 [0.34,0.43]
relevance	0.453	0.478	0.713	0.821	0.060	0.085	0.42 [0.37,0.46]	0.33 [0.29,0.37]

Jigsaw and IRep: stress tests. On the Jigsaw Specialized Rater Pools (Goyal et al., 2022; Google Jigsaw, 2022) (three ordinal pools, eight raters shared across pools) the test flags the LGBTQ and Control pools but not the African American pool. Sharing between pools biases $\hat{\tau}^2$ downward, so a flag under sharing is conservative (Appendix E). The covariance check singles out the LGBTQ–Control pair, which shares no rater. On IRep it flags Kuala Lumpur on most labels (25 of 31), and Budapest and Mexico City on fewer (15 and 8). Both datasets depart from the ideal design, so we report them qualitatively (Appendices G and B).

5 Discussion and limitations

The problem arises wherever an aggregate of human ratings is consumed, not only for summaries but also for RLHF preference data. The diagnostic assumes two independent panels. It is conservative under residual additive dependence, because shared variance is counted as signal, and a third panel exposes such dependence (Prop. 4; Appendix E; Algorithm 6). A good panel paired with one that measures something else also gets a low ceiling. Noise alone in the other panel does not lower it. With a third panel, the three pairwise covariances can help show which of the two is off (Appendix G). Either way, one panel's reliability number cannot show its fidelity.

The test needs three things: rater-level votes, two independent panels on the same items, and at least two raters per panel. Most references lack them, and those that have them are self-selected. A flag shows residual panel-specific variation under the model. It does not show that the panel measures an illegitimate construct. Where the contrast is a population, as in DICES and D3code, a flagged component is a difference to report, not an error to remove. Our judge results use only released outputs, so the judge set is what others published, and the nine-judge comparison rests on 200 sampled summaries rather than all $1{,}600$.

6 Conclusion

Two independent panels on the same items identify the common signal: what the panels share, which need not be ground-truth quality. They support a calibrated test of the panel-specific component. They also give a validity ceiling, under the additive model and for a judge whose errors are independent of the reference's. On SummEval, an LLM judge agrees with the expert panel more than even the upper confidence bound of that ceiling allows: its agreement with the experts includes structure that the independent crowd panel does not confirm. The certificate can also order references backwards. In 19 of the 28 D3code region pairs (11 with per-pair intervals excluding zero, 8 after a Bonferroni correction), the panel with the higher certificate is the less faithful one. On the Jigsaw pools, the point-estimate ranking by ρ_k is the reverse of the ranking by $\rho^\star$, with two of the three pairwise reversals significant. Across releases, panels share from 3% to 93% of their reliable variance, and the reported certificate does not tell which. Adding raters pushes the reliability number toward 1, while fidelity stays below a fixed cap. Most released references we surveyed cannot be checked. Yet making a reference checkable costs little: 1.1 percentage points of fidelity at 24 ratings per item, and $O(1/b)$ in general. Our rule: use two or more independent panels with replication, report the quantity consumed, and put the dependence test beside any ranking (Algorithm 1).

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to draft sections of the paper, for retrieval and discovery of related work, and for research ideation and execution, including writing and running the analysis code. We have reviewed all AI-assisted work: all theoretical claims and proofs were checked by the authors, the reported empirical figures were recomputed from the public rater-level releases and released judge outputs, and the cited figures on which our results rest (e.g. G-Eval's correlations) were checked against their sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyses only publicly released, de-identified human-annotation datasets (SummEval, DICES-990, D3code, MultiPref, the Jigsaw Specialized Rater Pools, and IRep), and released judge outputs (G-Eval; Guerdan et al., 2025; MultiPref), under the terms stated by their releases (MultiPref under ODC-BY); IRep is publicly released by its authors without an explicit licence, and we analyse it as released without redistributing it. The work involves no new data collection with human subjects. Our method is a measurement audit of existing references: the panel-specific component it estimates is a statistical quantity relative to a stated common-signal target, not a judgement of any rater group. Several of these references (DICES-990, D3code, IRep, Jigsaw) contrast rater populations by locale, culture, or identity; because treating one panel as ground truth can misattribute such construct or population differences as error, we report a detected panel-specific component as a population or cultural difference to be stated, never as a deficiency of any group, and give design diagnostics rather than certifications of any dataset or annotator pool.

REPRODUCIBILITY STATEMENT

All results are computed from public rater-level releases, released judge outputs, or simulations whose settings are fully stated in this manuscript. Appendix K lists, for every real-data analysis, the data source, preprocessing, coding, estimator and resampling scheme (Table 9), and, for every synthetic result, the model parameters and replications (Table 10); estimators are defined in §3.2 and Algorithms 1–6, and proofs are in Appendix A. All randomness uses NumPy `default_rng(0)`; point estimates are deterministic; bootstrap intervals are exact for our draw order, and simulated false-positive rates and powers carry Monte Carlo error. The metric correlations of Appendix C and G-Eval's published correlations are taken from Table 1 of Liu et al. (2023); the judge correlations with both panels (Appendix I) are recomputed from released judge outputs. The census of Appendix M applies a stated rule to released schemas. No separate supplementary material is submitted.

REFERENCES

Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. DICES dataset: Diversity in conversational AI evaluation for safety. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023. URL https://arxiv.org/abs/2306.11247.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081, 2025. doi: 10.18653/v1/2025.acl-long.782.

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959. doi: 10.1037/h0046016.

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M. Suchanek. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, 2022. URL https://arxiv.org/abs/2208.11646.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972.

Senjuti Dutta, Sid Mittal, Sherol Chen, Deepak Ramachandran, Ravi Rajakumar, Ian Kivlichan, Sunny Mak, Alena Butryna, and Praveen Paritosh. Modeling subjectivity (by mimicking annotator annotation) in toxic comment identification across diverse communities. arXiv preprint arXiv:2311.00203, 2023. URL https://arxiv.org/abs/2311.00203.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373.

Google Jigsaw. Specialized rater pools dataset. Kaggle, https://www.kaggle.com/datasets/google/jigsaw-specialized-rater-pools-dataset, 2022.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022. doi: 10.1145/3555088.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. Validating LLM-as-a-judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.05965.

Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24986–25004, 2025. doi: 10.18653/v1/2025.findings-emnlp.1361.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2506.07962.

G. Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. arXiv preprint arXiv:2605.29800, 2026. URL https://arxiv.org/abs/2605.29800.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, 2021. URL https://arxiv.org/abs/2106.04511.

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*, 2023. URL https://arxiv.org/abs/2303.16634.

Lester James V. Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL https://arxiv.org/abs/2410.19133.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18511–18526, 2024. URL https://aclanthology.org/2024.emnlp-main.1029/.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://arxiv.org/abs/2010.03532.

Jiaxin Pei and David Jurgens. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, 2023. URL https://arxiv.org/abs/2306.06826.

Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2306.01943.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Veerendra Kumar Sunkavalli. A closed-form estimator and diagnostic battery for anchor–judge error correlation, under a single-common-factor model. arXiv preprint arXiv:2609.08826, 2026. URL https://arxiv.org/abs/2609.08826.

Ka Wong and Praveen Paritosh. k-rater reliability: The correct unit of reliability for aggregated human annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 378–384, 2022. doi: 10.18653/v1/2022.acl-short.42.

Ka Wong, Praveen Paritosh, and Lora Aroyo. Cross-replication reliability – an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7053–7065, 2021. doi: 10.18653/v1/2021.acl-long.548.

Z. Yang, Y. Hou, and X. Yang. When the judge changes, so does the measurement: Auditing LLM-as-judge reliability. arXiv preprint arXiv:2607.08535, 2026. URL https://arxiv.org/abs/2607.08535.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://arxiv.org/abs/2210.07197.

A PROOFS

Propositions 5 and 4 are stated in Appendices F and E; their proofs are collected here.

A.1 MODEL

For panels $p = 1, \dots, P$, raters $j = 1, \dots, k$, and items $i = 1, \dots, n$,

$$y_{p,i,j} = \mu_i + c_{p,i} + \varepsilon_{p,i,j},$$

with $\mathrm{Var}(\mu) = \sigma_s^2$, $\mathrm{Var}(c_p) = \tau_p^2$ (i.i.d. across items, independent across panels), and $\mathrm{Var}(\varepsilon_p) = \sigma_p^2$, all mutually independent. The panel mean $\bar{y}_{p,i} = \mu_i + c_{p,i} + \bar{\varepsilon}_{p,i}$ satisfies $\mathrm{Var}(\bar{y}_{p,i}) = \sigma_s^2 + \tau_p^2 + \sigma_p^2/k$ and, for $p \neq q$, $\mathrm{Cov}(\bar{y}_{p,i}, \bar{y}_{q,i}) = \sigma_s^2$.

Proof of Proposition 1 (within-panel non-identifiability). Intuition. A single panel observes only item means, which fold the signal μ and the shared component c into one quantity; separating them needs a measurement that does not share c.

From a single panel the balanced one-way ANOVA sufficient statistics have $\mathbb{E}[\mathrm{MS}_{\text{between}}] = \sigma_p^2 + k(\sigma_s^2 + \tau_p^2)$ and $\mathbb{E}[\mathrm{MS}_{\text{within}}] = \sigma_p^2$. These identify σ_p^2 and the sum $\sigma_s^2 + \tau_p^2$, but no function of them separates σ_s^2 from τ_p^2; hence the fidelity $\rho_p^\star = \sigma_s^2/(\sigma_s^2 + \tau_p^2 + \sigma_p^2/k)$ is not identified. A two-way (item×rater) model adds a rater main effect and an interaction, but $c_{p,i}$ is constant across the raters of item i, so it is absorbed by the item main effect and attributable to neither added term. Finally $\hat{\rho}_k = (\sigma_s^2 + \tau_p^2)/(\sigma_s^2 + \tau_p^2 + \sigma_p^2/k) \to 1$ as $k \to \infty$, while $\rho_p^\star \to \sigma_s^2/(\sigma_s^2 + \tau_p^2) < 1$ whenever $\tau_p^2 > 0$. □

Proof of Proposition 2 (two-panel identification). Intuition. Two independent panels share only the signal, so their cross-panel covariance equals the signal variance exactly.

For independent panels $\mathrm{Cov}(\bar{y}_p, \bar{y}_q) = \mathrm{Var}(\mu) = \sigma_s^2$, since $c_p \perp c_q$ and the noises are independent; this identifies σ_s^2 and hence $\rho_p^\star = \sigma_s^2/\mathrm{Var}(\bar{y}_p)$. Given σ_p^2 (estimable from within-panel replication, $k \geq 2$), $\tau_p^2 = \mathrm{Var}(\bar{y}_p) - \sigma_s^2 - \sigma_p^2/k$ is identified. For $P > 2$ each of the $\binom{P}{2}$ pairs gives an unbiased estimate of σ_s^2; averaging reduces variance. □

Proof of Proposition 5 (detection variance). Intuition. The contrast W annihilates the signal part of the cross-panel covariance, leaving a variance that depends only on the non-signal term $g = \tau^2 + \sigma^2/k$ and the number of panels P.

Let $x_i = (M_{1,i}, \dots, M_{P,i})$ be the per-item vector of panel means, i.i.d. across items with covariance $\Sigma = \sigma_s^2 J_P + g\, I_P$, where $g = \tau^2 + \sigma^2/k$, J_P is all-ones and I_P the identity, since $\mathrm{Var}(M_{p,i}) = \sigma_s^2 + g$ and $\mathrm{Cov}(M_{p,i}, M_{q,i}) = \sigma_s^2$ for $p \neq q$. The leading term of the estimator is $\hat{D} = \sum_{a,b} W_{ab} \hat{\Sigma}_{ab}$ with the contrast $W = \frac{1}{P-1} I_P - \frac{1}{P(P-1)} J_P$ (diagonal $1/P$, off-diagonal $-1/(P(P-1))$), so $\mathbb{E}[\hat{D}] = \frac{1}{P} \sum_a (\sigma_s^2 + g) - \frac{1}{P(P-1)} \sum_{a \neq b} \sigma_s^2 = g$.

For Gaussian data, $\mathrm{Cov}(\hat{\Sigma}_{ab}, \hat{\Sigma}_{cd}) = \frac{1}{n}(\Sigma_{ac}\Sigma_{bd} + \Sigma_{ad}\Sigma_{bc}) + o(1/n)$, hence $\mathrm{Var}(\hat{D}) = \frac{2}{n} \sum_{abcd} W_{ab} W_{cd} \Sigma_{ac} \Sigma_{bd} = \frac{2}{n} \mathrm{tr}\big((W\Sigma)^2\big)$. Writing $W = \alpha I_P + \beta J_P$ with $\alpha = \frac{1}{P-1}$, $\beta = -\frac{1}{P(P-1)}$, and using $J_P^2 = P J_P$, $W J_P = (\alpha + \beta P) J_P = 0$; therefore $W\Sigma = W(\sigma_s^2 J_P + g I_P) = g\, W$. A direct computation gives $\mathrm{tr}(W^2) = P[\alpha^2 + 2\alpha\beta + \beta^2 P] = 1/(P-1)$, so $\mathrm{tr}((W\Sigma)^2) = g^2/(P-1)$ and $\mathrm{Var}(\hat{D}) = 2g^2/(n(P-1))$. When σ^2 is known, or estimated from an independent sample so that $\mathrm{Var}(\hat{\sigma}^2/k) = o(1/n)$ relative to this term, $\hat{\tau}^2 = \hat{D} - \sigma^2/k$ has the same leading variance, $\mathrm{Var}(\hat{\tau}^2) = 2(\tau^2 + \sigma^2/k)^2/(n(P-1))$. □

Proof of Proposition 3 (detectability excludes the fidelity optimum). Intuition. Detecting shared error needs within-panel replication ($k \geq 2$), whereas fidelity is maximised by spending the whole budget on single-rater panels ($k = 1$).

At $k = 1$ each panel gives one rating per item, so $c_{p,i}$ and $\varepsilon_{p,i,1}$ form a single draw and enter only through their sum; neither is identified for any P, so $\hat{\tau}_p^2$ is undefined without within-panel replication. Detection therefore requires $k \geq 2$, i.e. $P \leq b/2$. Meanwhile $R(P,k) = \sigma_s^2/(\sigma_s^2 + \tau^2/P + \sigma^2/b)$ is strictly increasing in P at fixed $b = Pk$, so its maximum is at $P = b$, $k = 1$, which lies outside the detectable region. For the price, $R(b,1) = \sigma_s^2/D_1$ and $R(b/2,2) = \sigma_s^2/D_2$ with D_1 and $D_2 = D_1 + \tau^2/b$ as in §3.3, so $R(b,1) - R(b/2,2) = \sigma_s^2(D_2 - D_1)/(D_1 D_2) = \sigma_s^2(\tau^2/b)/(D_1 D_2)$; since $D_1, D_2 = \sigma_s^2 + O(1/b)$, this is $\tau^2/(\sigma_s^2 b) + O(1/b^2)$. □

Proof of Proposition 4 (conservativeness and testability). Intuition. Whatever two panels share is counted as signal, so within the additive, nonnegative model residual dependence can only hide the panel-specific component, never invent it; a third panel exposes a sharing pair through an elevated pairwise covariance.

If panels p, q share a per-item component of variance $\gamma \geq 0$, then $\mathrm{Cov}(\bar{y}_p, \bar{y}_q) = \sigma_s^2 + \gamma$, so $\mathbb{E}[\hat{\tau}_p^2] = \mathrm{Var}(\bar{y}_p) - (\sigma_s^2 + \gamma) - \sigma_p^2/k_p = \tau_p^2 + \sigma_p^2/k_p - \gamma - \sigma_p^2/k_p = \tau_p^2 - \gamma \leq \tau_p^2$: the shared component is counted as signal and the estimator can only under-report; this conservativeness holds within the additive, nonnegative shared-component model and not for arbitrary negative or nonlinear dependence. Under mutual independence every pairwise cross-panel covariance equals σ_s^2; if a single pair shares $\gamma > 0$, that pair's covariance is elevated by γ while the others remain σ_s^2, so with $P \geq 3$ an unequal pairwise pattern identifies the dependent pair. A component shared by *all* panels shifts every covariance equally and is absorbed into σ_s^2, hence unidentifiable without external information. In simulation with three panels where two share $\gamma = 0.3$, the shared pair's covariance rises to 1.30 against 1.00 for the independent pairs. □

B IREP: PER-CLASS DETAILS

Cross-replication reliability (Wong et al., 2021) is the closest prior work, so we also apply the test to the dataset it released. In IRep, three rater pools (Budapest, Mexico City, Kuala Lumpur; the release's International pool is excluded) label the same items, with up to two raters per pool. As in Wong et al. (2021), we analyse 31 labels (30 categories plus "unsure"). The released CSV holds $127{,}078$ rater–item records on $38{,}499$ items for the three city pools, that is $3{,}939{,}418$ label judgements, the figure Wong et al. report. For each emotion class c we set $y_{p,i,j}^{(c)} = \mathbf{1}[\text{rater } j \text{ in pool } p \text{ labels item } i \text{ with } c]$ and run the *same* estimator on these binary indicators. This gives one $\hat{\tau}^2$ and one interval per class. Classes are analysed separately, and the intervals are *not* corrected for multiplicity, so we read IRep qualitatively, as a stress test. The test again *discriminates*. We use the $4{,}248$ items rated by two raters in each city pool. For "interest" and "concentration" it flags the Kuala Lumpur pool ($\hat{\tau}^2 = 0.092$, 90% interval $[0.086, 0.097]$; and 0.085, $[0.080, 0.090]$) but not Budapest or Mexico City. For "contentment" all three pools flag. This matches Wong et al. (2021), who report internal reliabilities of 0.4494 (Mexico City) and 0.6363 (Kuala Lumpur) for that class against a cross-pool κ_x of -0.0344. The three pairwise covariances are uneven, so the covariance-consistency check (Prop. 4) shows which pools share the component. Two caveats apply. The categorical construct requires this per-class reading, and with only two raters per pool ($k = 2$) the within-pool variance is noisy. The point is qualitative. Applied to xRR's own data, the covariance-based diagnostic finds panel-specific structure that a descriptive agreement coefficient does not identify.

C PER-METRIC NOISE-CEILING BREAKDOWN

The pooled expert fidelities of Table 2 have wide 90% item-bootstrap intervals: $[-0.00, 0.06]$ for coherence, $[-0.03, 0.03]$ for consistency, $[0.01, 0.09]$ for fluency and $[0.02, 0.09]$ for relevance. Table 3 takes each metric's summary-level Spearman correlation with the SummEval expert annotations from Table 1 of Liu et al. (2023). These values follow the same protocol as G-Eval and are cited, not recomputed. The ceilings, by contrast, are our own computations from the SummEval ratings, on document-centred panel means with a document bootstrap; they are not taken from prior work. We first computed the ceilings on pooled item means; that version overstates them, because summary-level correlations never see differences between documents. We express each one as a per-

centage of the within-document ceiling $\sqrt{\rho_k^E}$ (bottom row), so that entries can be compared across dimensions.

Table 3: **[Real data]** SummEval: summary-level Spearman correlation with the expert annotations (Liu et al., 2023, Table 1), as % of the within-document ceiling $\sqrt{\rho_k^E}$ (bottom row; §4).

% of ceiling	coherence	consistency	fluency	relevance
ROUGE-1	19	16	12	40
ROUGE-2	21	19	17	36
ROUGE-L	15	12	11	39
BERTScore	32	11	21	39
MoverScore	18	16	14	39
BARTScore	51	39	38	44
UniEval	65	45	48	53
G-Eval-4	66	52	48	68
ceiling $\sqrt{\rho_k^E}$ (within-document)	0.879	0.982	0.941	0.806

G-Eval attains 48–68% of the ceiling. It is highest or tied-highest in every dimension (tied with UniEval on fluency): closest on relevance (68%) and coherence (66%), furthest on fluency (48%). The overlap and embedding metrics (ROUGE, BERTScore, MoverScore) reach only 11–40%. Because the ceiling varies sharply by dimension, the same raw correlation means different things in different dimensions. That is why we report ceiling-normalised figures.

Sensitivity of the G-Eval result. Table 4 recomputes G-Eval's correlation with each panel in four ways, with the scoring and join of Appendix I. The split holds in all four: 0.47–0.61 with the experts and 0.02–0.09 with the crowd. The Pearson rows matter most, because there the attenuation bound is exact. On pooled summaries, the upper bounds of the pooled expert fidelity above are 0.06, 0.03, 0.09 and 0.09. Their square roots, the validity ceilings, are about 0.3 at most. G-Eval's pooled Pearson correlation with the experts (0.58–0.61) is above them in every dimension.

Table 4: Recomputed G-Eval: correlation with the expert / crowd panel mean under four choices of correlation and aggregation (point estimates).

dimension	per-document Spearman	per-document Pearson	pooled Spearman	pooled Pearson
coherence	0.571 / 0.021	0.585 / 0.035	0.572 / 0.022	0.581 / 0.040
consistency	0.500 / 0.025	0.591 / 0.038	0.504 / 0.038	0.575 / 0.037
fluency	0.506 / 0.040	0.593 / 0.040	0.470 / 0.027	0.600 / 0.024
relevance	0.564 / 0.057	0.588 / 0.067	0.590 / 0.083	0.613 / 0.090

Algorithm 2 AUDITREFERENCE$(Y_1, Y_2;\ B = 500,\ \alpha = 0.10)$: two-panel audit

Input: rating matrices Y_1, Y_2 (items $\times$ raters, same items in the same order); bootstrap draws B; level α. Conventions (ddof, unequal k) as in Appendix K.

1. Items. Keep items with at least two ratings in each panel. If there are none, return NON-IDENTIFIABLE UNDER THE OBSERVED DESIGN.

2. One-way ANOVA per panel. Item means $\bar{y}_p$, within-item variance $\hat{\sigma}_p^2$, mean $\bar{\kappa}_p$ of $1/k_{p,i}$, and between-item variance $V_p = \mathrm{Var}(\bar{y}_p)$.

3. Common signal. $c = \mathrm{Cov}(\bar{y}_1, \bar{y}_2)$, the estimate of σ_s^2. With three or more panels, use for panel p the mean of its pairwise covariances.

4. Per panel. $\rho_{k,p} = 1 - \hat{\sigma}_p^2 \bar{\kappa}_p / V_p$; fidelity $\rho_p^\star = c/V_p$; validity ceiling $\sqrt{\max(\rho_p^\star, 0)}$; component $\hat{\tau}_p^2 = V_p - c - \hat{\sigma}_p^2 \bar{\kappa}_p$; shared fraction $c/(c + \max(\hat{\tau}_p^2, 0))$; cross-panel $r = c/\sqrt{V_1 V_2}$.

5. Percentile bootstrap over items. Repeat steps 2–4 on B resamples; take the $\alpha/2$ and $1 - \alpha/2$ percentiles of each quantity.

6. Verdicts. Panel p is FLAGGED if the lower percentile of $\hat{\tau}_p^2$ exceeds 0, IDENTIFIED otherwise. A certificate–fidelity inversion holds when $\rho_{k,1} - \rho_{k,2}$ and $\rho_1^\star - \rho_2^\star$ have opposite signs; it is significant when both intervals exclude 0.

Output: c, r; per panel $\rho_{k,p}$, $\rho_p^\star$, validity ceiling, $\hat{\tau}_p^2$ with interval and flag, shared fraction; inversion; verdict.

Algorithm 3 JUDGECHECK$(J, Y_1, Y_2;\ B)$: a judge against both panels

Input: judge scores J_i on the same items; rating matrices Y_1, Y_2; B.

1. Correlations. $\mathrm{Corr}(J, \bar{y}_1)$ and $\mathrm{Corr}(J, \bar{y}_2)$, with the correlation the benchmark reports (Pearson, or per-document Spearman averaged over documents; Appendix I).

2. Percentile bootstrap over items (or documents): intervals for both correlations and for their gap.

3. Ceilings. Run AUDITREFERENCE(Y_1, Y_2), on document-centred panel means when the correlations are per-document (Appendix K); read each panel's validity ceiling and its upper percentile.

4. Verdict. PANEL FITTING toward panel p if $\mathrm{Corr}(J, \bar{y}_p)$ exceeds the upper percentile of panel p's ceiling, or if the gap interval excludes 0 in favour of p while $\mathrm{Corr}(J, \bar{y}_q)$ is not significantly positive; CONFIRMED BY BOTH PANELS if both correlation intervals lie above 0 and neither correlation exceeds its ceiling; UNRESOLVED otherwise.

Output: both correlations, the gap and its interval, both ceilings, the verdict. The comparison assumes the additive model and judge errors independent of the reference's; exact for Pearson, approximate for Spearman.

D THE AUDIT PROCEDURE

Algorithm 1 Reliability audit of a human reference

Input: rater-level vote matrix; the aggregation size k the benchmark consumes; optionally a second independent panel on the same items.

1. Unit and figure. Estimate ρ_1 and ρ_k by one-way ANOVA; report which the benchmark consumes and $k_{\min} = \lceil \rho_k^{\mathrm{tgt}}(1 - \rho_1)/(\rho_1(1 - \rho_k^{\mathrm{tgt}})) \rceil$ for a target ρ_k^{tgt}.

2. Identifiability. From the incidence structure: if it cannot support the requested decomposition (e.g. no within-panel replication), return NON-IDENTIFIABLE UNDER THE OBSERVED DESIGN.

3. Dependence (requires ≥ 2 panels, $k \geq 2$). Compute $\hat{g}_p = \widehat{\mathrm{Var}}(\bar{y}_p) - \widehat{\mathrm{Cov}}(\bar{y}_p, \bar{y}_q)$ and $\hat{\tau}_p^2 = \hat{g}_p - \hat{\sigma}_p^2/k$; take an item-bootstrap one-sided interval and flag when its lower bound exceeds 0. With ≥ 3 panels, compare the pairwise cross-panel covariances as a covariance-consistency check (an unequal pair flags dependence; equality does not prove independence).

4. Report two verdicts. *Reference-audit status*: IDENTIFIED / FLAGGED (panel-specific shared component detected) / NON-IDENTIFIABLE UNDER THE OBSERVED DESIGN. *Ranking status*: bring compared figures to the same unit and return RESOLVED (with the ranking) if the reliability-aware interval separates the panels, else UNRESOLVED; a FLAGGED reference cautions interpretation rather than forcing abstention.

Output: the three declarations, the reference-audit status, and the ranking status.

Algorithms 2–6 restate the audit and its design rules as functions, with their inputs, outputs and every computation, so that it can be implemented directly from this appendix.

Algorithm 4 DESIGNPANELS$(b,\ \lambda,\ n)$: choosing panels under a fixed budget

Input: budget b ratings per item; a planning value of $\lambda = \tau^2/\sigma^2$ (from a pilot, or a range); number of items n. Variances are in units of σ^2, with σ_s^2 set to its planning value (1 if unknown).

1. Feasible designs. List the pairs (P, k) with $Pk = b$, $P \geq 2$ and $k \geq 2$. If there are none ($b < 4$), return NON-IDENTIFIABLE UNDER THE OBSERVED DESIGN: the budget cannot buy two panels with replication.

2. Per-panel flags. If each panel will be audited on its own (Algorithm 2), recommend P between 2 and 4 among the feasible designs; detection is nearly flat over this range (Appendix F).

3. Pooled estimate. If a common τ^2 is to be estimated, compute for each feasible (P, k) the SNR $\lambda\sigma^2/\sqrt{2(\lambda\sigma^2 + \sigma^2/k)^2/(n(P-1))}$ (Prop. 5) and return the feasible P that maximises it (exhaustive search; no closed form is assumed).

4. Price of auditability. Report $R(b, 1) - \max_{\text{feasible}} R(P, k)$, with $R(P, k) = \sigma_s^2/(\sigma_s^2 + \tau^2/P + \sigma^2/b)$: the fidelity given up by keeping the design checkable (Prop. 3).

Output: the feasible designs, P and k for per-panel and pooled use, and the price of auditability.

Algorithm 5 ITEMSNEEDED$(Y_{\text{pilot}},\ \gamma,\ \alpha,\ \beta)$: how many items to annotate

Input: pilot rating matrices for two or more panels on n_0 shared items, with $k \geq 2$; target shared-error fraction $\gamma = \tau_p^2/\sigma_s^2$; level α; power $1 - \beta$; bootstrap draws B.

1. Pilot estimates. Run steps 2–3 of Algorithm 2 on the pilot to get $\hat{\sigma}_s^2 = c$.

2. Per-item variance. Bootstrap the pilot items B times, recompute $\hat{\tau}_p^2$ on each draw, and set $\widehat{\text{Var}}_1 = n_0 \cdot \widehat{\text{Var}}_{\text{boot}}(\hat{\tau}_p^2)$.

3. Required items. Return $n = \lceil (z_{1-\alpha} + z_{1-\beta})^2 \, \widehat{\text{Var}}_1/(\gamma\hat{\sigma}_s^2)^2 \rceil$ (Appendix F). Because $\widehat{\text{Var}}_1$ comes from a small pilot, round n up.

Output: the number of items n, and the predicted power at any other n: $\Phi\big(\gamma\hat{\sigma}_s^2/\sqrt{\widehat{\text{Var}}_1/n} - z_{1-\alpha}\big)$.

Algorithm 6 COVARIANCECHECK$(Y_1, \ldots, Y_P;\ B,\ \alpha)$: are three or more panels independent?

Input: rating matrices for $P \geq 3$ panels on the same items; B; level α.

1. Pairwise covariances. Compute the panel means and the $\binom{P}{2}$ covariances $c_{pq} = \widehat{\text{Cov}}(\bar{y}_p, \bar{y}_q)$ (ddof $= 1$).

2. Contrast. For each pair, $d_{pq} = c_{pq} -$ the mean of the other pairwise covariances.

3. Bootstrap. Draw items with replacement B times and recompute every d_{pq}. With $m = \binom{P}{2}$ pairs, take for each pair the lower $\alpha/(2m)$ percentile (Bonferroni over pairs).

4. Verdict. A pair is ELEVATED when that lower percentile exceeds 0: its panels share a component beyond the common signal, which Prop. 4 counts as signal for that pair. If no pair is elevated, return NO DETECTED DEPENDENCE; this does not prove independence, and a component shared by *all* panels stays undetectable.

Output: the pairwise covariances, the elevated pairs, and the verdict.

Table 5: Audit card for a human reference, with two examples from this paper (SummEval coherence: Tables 2 and 1, Appendix I; DICES-990: Appendix H).

field	SummEval, coherence	DICES-990
panels and k	experts $k = 3$; crowd $k = 5$	US; India (≈ 36 ratings per item each)
ρ_1 (panel 1 / panel 2)	0.559 / 0.494	–
ρ_k (panel 1 / panel 2)	0.792 / 0.830	0.90 / 0.88
cross-panel r	0.04	0.83
$\rho^\star$ (panel 1 / panel 2)	0.027 / 0.058	0.84 / 0.81
$\hat{\tau}^2$ (panel 1 / panel 2)	0.82 / 0.39, both flagged	≈ 0.002, both flagged
shared fraction (flagged panel)	3%	93%
judge vs. panel 1 / panel 2	G-Eval 0.571 / 0.021: panel fitting	–

Using the audit in practice. The algorithms above cover five situations.

1. *Planning a new reference.* Choose the number of panels and raters with Algorithm 4, and the number of items from a small pilot with Algorithm 5. Keep at least two raters in each of at least two independent panels, so that the reference can be checked.
2. *Auditing a reference with two panels.* Run Algorithm 2. Report ρ_k and $\rho^\star$ for each panel, and $\hat{\tau}^2$ with its interval and flag. When the panels contrast populations, report a flagged component as a population difference, not as annotator error.
3. *Three or more panels.* Run Algorithm 6 first. An elevated pair shares a component beyond the common signal; for that pair it is counted as signal (Prop. 4). Then run Algorithm 2, using for each panel the mean of its pairwise covariances.
4. *Validating an LLM judge.* Run Algorithm 3. Report the judge's correlation with each panel, the gap between them, and each panel's validity ceiling. A judge that agrees with one panel and not the other is fitting that panel.
5. *A single panel only.* Report ρ_1, ρ_k, the k the benchmark consumes and $k_{\min}$ (Algorithm 1, step 1), and state that fidelity to a common signal cannot be checked under this design (Prop. 1).

An audit card. Table 5 lists what a reference should report, filled in with two references from this paper. A dash marks a quantity we do not report for that reference.

Checks of Algorithms 4–6. We ran each of them on the settings of this paper (Table 10). DESIGNPANELS at $b = 24$ returns the pooled optima of Fig. 2 and the 1.1-point price. ITEMSNEEDED, from 200-item pilots of the calibration setting, asks for 135 items for 80% power; at that size the simulated power is 84%, so it errs on the safe side. COVARIANCECHECK flags only the LGBTQ–Control pair on Jigsaw, as in Appendix G. In simulation, with two of three panels sharing a component, it flags that pair 94% of the time and never an independent pair; when all three panels are independent, it flags some pair 6.5% of the time.

E ROBUSTNESS TO THE INTER-PANEL INDEPENDENCE ASSUMPTION

The test assumes that the panels' shared errors are independent. If the audit is applied to panels of LLM judges rather than human raters, this calls for panels from diverse model families. Two facts address the assumption.

Proposition 4 (Conservative within the additive model; checkable with three panels)**.** *(i) Within the assumed additive, nonnegative shared-component model, if two panels share a per-item component of variance $\gamma \geq 0$ beyond the signal, then $\mathrm{Cov}(\bar{y}_p, \bar{y}_q) = \sigma_s^2 + \gamma$, so $\hat{\tau}_p^2 = \hat{g}_p - \hat{\sigma}_p^2/k_p$ has expectation $\tau_p^2 - \gamma$, omitting that shared component: the diagnostic counts the shared variance as signal and can only* under*-report the panel-specific component, never manufacture it. This conservativeness is a property of that model; arbitrary negative or nonlinear residual dependence is outside the guarantee. (ii) With $P \geq 3$ panels, mutual independence implies all pairwise cross-panel covariances*

equal σ_s^2; comparing them therefore provides a covariance-consistency diagnostic for deviations from the assumed structure, and an elevated pair identifies which panels share error. Equal pairwise covariances do not establish independence or exclude higher-order or nonlinear dependence.

Simulation confirms both (Appendix A). The remaining limit is fundamental and not specific to this test. A component shared by *all* panels is absorbed into σ_s^2 and cannot be identified without external ground truth. So "signal" always means "what every panel agrees on."

F THE DETECTION-OPTIMAL ALLOCATION

Proposition 5 (Detection variance)**.** *For P balanced panels of k raters on n items, with a shared-error variance τ^2 common to all panels and σ^2 known or pre-estimated, the pooled shared-error estimator obeys, to leading order in n,*

$$\mathrm{Var}(\hat{\tau}^2) = \frac{m\,(\tau^2 + \sigma^2/k)^2}{n\,(P-1)}, \qquad m = 2 \text{ under Gaussianity.}$$

The constant m is not a design parameter but a distributional one: it is the Gaussian fourth-moment value $m = 2$, and for heavier-tailed ratings it becomes $m = 2 + \bar{\kappa}$ with $\bar{\kappa}$ an excess-kurtosis term, so heavier tails inflate the variance and reduce detection power (more items are then needed). The test does not rely on m: it uses the item bootstrap; m enters only this analytic approximation.

The proof (App. A) uses a contrast W that removes the signal part of the cross-panel covariance, and reduces to $\mathrm{tr}((W\Sigma)^2) = (\tau^2 + \sigma^2/k)^2/(P-1)$. In simulation, the $m = 2$ form is accurate to within 2–6% (0.98, 1.03 and 1.06 times the formula).

Panel-sizing. Inverting Spearman–Brown, the smallest panel reaching a target aggregate reliability ρ_k^{tgt} from a single-rating ρ_1 is $k_{\min} = \left\lceil \frac{\rho_k^{\mathrm{tgt}}(1-\rho_1)}{\rho_1(1-\rho_k^{\mathrm{tgt}})} \right\rceil$, which we recommend reporting instead of comparing ρ_1 to a fixed threshold.

Items and power. The standard error of $\hat{\tau}_p^2$ scales as $O(n^{-1/2})$. To detect a shared-error fraction $\tau_p^2/\sigma_s^2 = \gamma$ at level α with power $1-\beta$, one needs $n \gtrsim (z_{1-\alpha} + z_{1-\beta})^2\,\mathrm{Var}_1/(\gamma\sigma_s^2)^2$ items. Here Var_1 is the per-item influence variance, estimated by the same bootstrap.

Per-panel estimator. Prop. 5 and the figure concern the pooled estimator of a τ^2 common to all panels. The test flags panels one at a time, with $\hat{\tau}_p^2 = \widehat{\mathrm{Var}}(\bar{y}_p) - \widehat{\mathrm{Cov}}(\bar{y}_p, \bar{y}_q) - \hat{\sigma}_p^2/k$, the covariance averaged over the other panels. Its variance is larger and depends on σ_s^2. In simulation at $b = 24$ ($n = 200$, $\sigma_s^2 = \sigma^2 = 1$, 1,500 replications) its detection SNR peaks at $P = 3$ for $\lambda \in \{0.3, 0.5, 1.0\}$ and at $P = 2$ for $\lambda = 0.1$, and it is nearly flat between $P = 2$ and $P = 4$. A second simulation run gives $P = 4$ at $\lambda = 1$; the choice among two to four panels matters little. The pooled optimum, by contrast, grows with λ: $P = 4$, 8, 12, 12 with σ^2 known (Fig. 2), and $P = 4$, 6, 8, 12 with σ^2 estimated, as in this simulation. Neither is at $k = 1$.

Integer optimum (pooled estimator). We minimise the leading-order $\mathrm{Var}(\hat{\tau}^2)$, which is the same as maximising detection SNR, over the feasible integers $P \geq 2$, $k = b/P \geq 2$. With σ^2 *known*, the unconstrained analytic minimiser is $P^\star_{\mathrm{dom}} = \lambda b + 2$. Within the feasible region $2 \leq P \leq b/2$ this is an interior optimum when $\lambda b + 2 \leq b/2$; otherwise the optimum sits at the boundary $P = b/2$ ($k = 2$). Exhaustive integer search confirms this. When σ^2 is estimated jointly, its estimation cost grows with P and moves the optimum to somewhat fewer panels, as simulation confirms. We therefore report the exact integer comparison and do not claim a universal closed form. The robust law is qualitative. The detection optimum grows with λ: it is interior for small λ, at the boundary $P = b/2$ for large λ (Fig. 2), and never at the fidelity optimum $k = 1$. With a fixed cost per panel the budget becomes $C = n(c_P P + c_R P k)$, and a large c_P pulls the detection optimum toward fewer, larger panels. The mismatch persists.

G JIGSAW AND IREP: FULL RESULTS

Jigsaw. We run the test on the Jigsaw Specialized Rater Pools (Goyal et al., 2022), released on Kaggle (Google Jigsaw, 2022). Three rater pools (African American, LGBTQ, Control) each rate the same $25{,}500$ comments, with five raters per pool, on the toxicity scale (-2 very toxic, -1 toxic, 0 unsure, 1 not toxic), which we treat as continuous. We drop null ratings (comment marked incomprehensible), which leaves $25{,}490$ comments with at least two ratings in every pool. We take each pool's common-signal covariance as the mean of its two pairwise covariances. The test then flags LGBTQ ($\hat{\tau}^2 = 0.040$, 90% interval $[0.036, 0.045]$) and Control (0.022, $[0.017, 0.026]$), but not African American (-0.015, $[-0.018, -0.011]$). The pairwise covariances are 0.210 (African American–LGBTQ), 0.210 (African American–Control) and 0.240 (LGBTQ–Control). The check singles out LGBTQ–Control, elevated by about 14%, although these two pools share no rater. (The release's pseudonymous IDs place 5 raters in both African American and LGBTQ, and 3 in both African American and Control.) The elevation therefore reflects structure common to those two pools beyond rater identity. Prop. 4 reads it as cross-pool sharing γ and counts it as signal for that pair. Neither of African American's covariances is elevated (0.210, 0.210), so cross-pool sharing cannot explain its interval entirely below zero. A pool that tracks the common signal with a smaller loading ($\ell_p < 1$) produces exactly this, since then $\mathbb{E}[\hat{\tau}_p^2] = \tau_p^2 - \ell_p(1-\ell_p)\sigma_s^2$. We read it as "no detectable panel-specific component" plus a scale difference, not as a negative variance. Because the pools share raters, we do not claim the nominal calibration for this dataset.

IRep. Per-class results are in Appendix B.

H DICES, D3CODE AND MULTIPREF: FULL RESULTS

DICES: a native case with two independent pools. In DICES-990 (Aroyo et al., 2023), US and India rater pools (80 and 92 raters; 29 to 45 ratings per item, about 36 on average) score the same 990 chatbot-safety conversations. This is the two-panel design, native to LLM evaluation and far more replicated than SummEval. The pools agree well: cross-panel $r = 0.83$, and 93% of each pool's reliable between-item variance is shared. Each aggregate therefore tracks the common signal ($\rho^\star \approx 0.84$ US and 0.81 India, close to $\rho_k \approx 0.90$ and 0.88). The test still flags a small but significant panel-specific component ($\hat{\tau}^2 \approx 0.002$, 90% lower bound above zero for both pools). This fits a US–India *population* difference in safety perception, and it holds whichever way "Unsure" is coded. DICES is the opposite of SummEval: high reliability that is *also* auditable, with a real but bounded population effect that the coefficient alone would hide.

D3code: a second dataset. In D3code (Mostafazadeh Davani et al., 2024), eight geo-cultural regions rate the same $4{,}554$ offensiveness items. Each region has a median of 3 ratings per item (range 1 to 28); each ρ_k uses the unequal-k correction of Appendix K. Regional pools agree only moderately (cross-panel $r = 0.26$–0.47 over the 28 region pairs), and the test *discriminates*. For North America vs. Sinosphere it flags a large North-American component ($\hat{\tau}^2 = 0.24$, 90% lower bound 0.20) but not the Sinosphere pool (-0.02). For Western Europe vs. Arab Culture it flags both (0.08 and 0.10, both lower bounds above zero). Over all 28 pairs, at least one pool is flagged in every pair, and both pools in 11. Only one of these 11 is among the 11 significant inversions of Appendix J. The pattern is thus not specific to one dataset. As with DICES, a flagged component is a cultural difference: a caveat on the reference's scope, not a defect.

MultiPref: preference data. MultiPref (Miranda et al., 2025) gives each of $10{,}461$ response comparisons two annotations from a non-expert pool and two from an expert pool. The 89 non-expert and 138 expert identifiers in the release are disjoint, so the two pools are independent panels on the same items. On overall preference (five-point scale) the pools agree moderately (cross-panel $r = 0.41$), and both aggregates are only modestly reliable at $k = 2$ ($\rho_k = 0.42$ for experts, 0.45 for non-experts). Yet almost all of that reliability is shared: fidelity to the common signal is $\rho^\star = 0.43$ and 0.40. The expert $\rho^\star$ slightly exceeds its ρ_k because $\hat{\tau}_E^2$ is negative (-0.005, an interval covering zero). Under the model $\rho^\star \le \rho_k$ exactly when $\tau_p^2 \ge 0$, so this inversion reflects sampling noise, not a component. The test flags a non-expert-specific component ($\hat{\tau}^2 = 0.052$, 90% interval $[0.025, 0.077]$) but not an expert-specific one (-0.005, $[-0.031, 0.021]$). Helpfulness shows the

same pattern. For truthfulness it is the expert pool that flags (0.015, $[0.005, 0.025]$), and for harmlessness neither pool does (0.001, and -0.004 with an interval entirely at or below zero). The audit here returns the opposite of SummEval: a *low* certificate that is honest about what it measures, with a small, dimension-specific panel component that the coefficient alone would not reveal.

I JUDGES AGAINST BOTH PANELS

G-Eval on SummEval. We recompute G-Eval from its released per-summary outputs: GPT-4, 20 sampled responses per summary, scored by the mean of the parsed numbers. The published scores weight by token probabilities instead, so our values differ slightly. We correlate the recomputed score, at the summary level, with each panel's mean on the same $1{,}600$ summaries.

Table 6: Recomputed G-Eval: summary-level Spearman correlation with each SummEval panel (90% document-bootstrap intervals).

dimension	vs. experts	vs. crowd	panel–panel r
coherence	0.571 $[0.539, 0.603]$	0.021 $[-0.023, 0.063]$	0.004
consistency	0.500 $[0.474, 0.528]$	0.025 $[-0.021, 0.073]$	-0.025
fluency	0.506 $[0.465, 0.545]$	0.040 $[-0.015, 0.091]$	0.058
relevance	0.564 $[0.523, 0.601]$	0.057 $[0.012, 0.102]$	0.030

The recomputed expert correlations are within 0.02 of the published 0.582, 0.507 and 0.547 for coherence, consistency and relevance. For fluency, which G-Eval scores on a 1–3 scale, we obtain 0.506 against a published 0.455. The expert–crowd gap in G-Eval's correlation excludes zero in every dimension (intervals $[0.49, 0.61]$, $[0.42, 0.53]$, $[0.40, 0.53]$, $[0.45, 0.57]$). The crowd correlation is indistinguishable from zero in three dimensions. G-Eval tracks the expert panel, and the two panels barely agree with each other.

Nine recent judges on SummEval. Guerdan et al. (2025) release the responses of nine LLM judges on 200 SummEval summaries per dimension, sampled by the authors to limit computation (`lguerdan/indeterminacy-experiments`, `runs/main-run/summ_eval_*`). Each judge answers a forced-choice question ten times (eight for o3-mini). The released table is indexed [item][forced-choice option][response set][draw]. The first option is the positive category (e.g. "Coherent"). We sum it over response sets and draws, and divide by all recorded answers. Null answers are rare: 50 of $6{,}400$ for o3-mini (0.8%, at most 2.1% in one dimension), 3 for Mistral Large, and none for the other seven judges. We match the items to SummEval by document identifier and summary text, and all 200 match. For the 7 items whose text matches several system records, we use the first one in file order. We then correlate the score with each panel's mean on the original 1–5 scale, using item-level Pearson correlation and $1{,}000$ item draws. Items without a valid answer are dropped (44 for Claude 3.5 Sonnet on coherence, 1 for Mistral Large; 200 otherwise). We also tried scoring each judge by its modal answer (positive when more than half of the answers choose the positive option). That gives a slightly smaller expert–crowd gap (mean 0.38 against 0.40, smaller in 28 of the 36 pairs), so we report the fraction.

Table 7: Nine LLM judges (Guerdan et al., 2025): Pearson correlation with the expert / crowd panel mean on 200 SummEval summaries. $^\circ$: the 90% interval of the expert–crowd gap covers zero (all other 33 exclude it).

judge	coherence	consistency	fluency	relevance
Claude 3.5 Sonnet	0.42 / -0.03	0.64 / 0.05	0.18 / 0.07	0.49 / 0.11
Claude 3 Haiku	0.38 / 0.00	0.21 / -0.07	0.20 / -0.05	0.38 / -0.11
DeepSeek Chat	0.52 / 0.02	0.66 / 0.05	0.33 / 0.08	0.54 / -0.03
Llama 3.3 70B	0.62 / 0.00	0.68 / 0.03	0.11 / 0.01	0.57 / -0.01
Mistral Large	0.53 / 0.01	0.62 / -0.01	0.09 / 0.03°	0.53 / 0.10
Mistral Small	0.38 / -0.03	0.63 / -0.02	0.08 / 0.10°	0.45 / 0.07
GPT-3.5 Turbo	0.36 / -0.04	0.21 / 0.03	0.19 / 0.02	0.37 / 0.02
GPT-4o mini	0.35 / -0.02	0.58 / 0.00	0.09 / 0.01°	0.50 / 0.02
o3-mini	0.44 / 0.01	0.65 / 0.01	0.51 / 0.02	0.54 / 0.13
panel–panel r	-0.03	0.06	0.01	0.13

In 33 of the 36 pairs, the judge tracks the expert panel far more than the crowd panel. Correlation with the crowd stays within $[-0.11, 0.13]$ throughout, while the two panels themselves barely agree on these items. One of the 33 is borderline: for Llama 3.3 on fluency the lower bound of the gap changes sign across bootstrap seeds. The three exceptions are fluency judges whose expert correlation is itself below 0.10.

GPT-4 on MultiPref. The release's `gpt4_overall_binarized` configuration gives a GPT-4 preference for each of the $10{,}461$ comparisons. Its Pearson correlation is 0.328 $[0.314, 0.344]$ with the expert panel mean and 0.369 $[0.356, 0.383]$ with the non-expert panel mean; the two panels correlate at 0.413. On non-tied items it agrees with the experts' direction 67.5% of the time and with the non-experts' 69.9%. Here the panels share almost all their reliable variance (Appendix H). The judge's agreement is confirmed by both, and it sits well below either validity ceiling ($\sqrt{0.43} = 0.66$, $\sqrt{0.40} = 0.63$).

J Certificate–fidelity reversals

Two panels p, q on the same items share one common-signal covariance c. So $\rho_p^\star = c/\mathrm{Var}(\bar{y}_p)$ and $\rho_p^\star/\rho_q^\star = \mathrm{Var}(\bar{y}_q)/\mathrm{Var}(\bar{y}_p)$, while panel p's ρ_k is $1 - \hat{\sigma}_p^2\,\overline{k_p^{-1}}/\mathrm{Var}(\bar{y}_p)$. A panel-specific component τ_p^2 enlarges $\mathrm{Var}(\bar{y}_p)$. That raises ρ_k and lowers $\rho^\star$, so the certificate prefers the panel that is *less* faithful to what the two panels share. For every panel pair we check whether the panel with the higher ρ_k has the lower $\rho^\star$ (a reversal). We call it significant when the 90% paired item-bootstrap intervals of $\rho_{k,p} - \rho_{k,q}$ and $\rho_p^\star - \rho_q^\star$ both exclude zero, with opposite signs (500 draws, `default_rng(0)`). Gaps use each pair's own covariance. So the Jigsaw Control–LGBTQ gap in Fig. 1d (-0.004) differs in the third decimal from the pool-level values below ($0.548 - 0.551 = -0.003$).

Table 8: Certificate–fidelity reversals by dataset.

dataset	panel pairs	reversals	significant
D3code (region pairs)	28	19	11
Jigsaw SRP (pool pairs)	3	3	2
MultiPref (four dimensions)	4	3	0
SummEval (four dimensions)	4	2	0
DICES-990	1	0	0

We also apply a Bonferroni correction over the 28 D3code pairs, at level $0.10/28$ per interval with $4{,}000$ paired draws. Then 8 of the 11 significant reversals remain, and all 8 involve Sinosphere or the Indian Cultural Sphere. Ten of the eleven significant D3code reversals involve Sinosphere or the Indian Cultural Sphere; the eleventh is North America vs. Oceania. On Jigsaw the order by ρ_k (LGBTQ 0.646, Control 0.604, African American 0.557) is the reverse of the order by $\rho^\star$ (African American 0.598, Control 0.551, LGBTQ 0.548). Both pairs involving African American are significant. The Control–LGBTQ pair differs by only 0.003 in fidelity and is not. In these cases, the reference chosen by its reliability certificate is the less faithful one.

K Parameters and constants for reproduction

Every reported figure can be regenerated from the public releases and the settings below. A second implementation written from this appendix reproduces every point estimate. All randomness uses NumPy `default_rng(0)` (PCG64, seed 0). Each result (each row of Tables 9 and 10) starts its own `default_rng(0)`; within a result the draws form one stream, except where a row says otherwise. Intervals are 90% percentile intervals, read one-sided at 5% for the $\hat{\tau}^2$ test. Point estimates are deterministic; bootstrap intervals are exact for our draw order and otherwise reproduce up to Monte Carlo resampling order. Within-item variances use $\mathrm{ddof} = 1$ and are pooled by averaging over items. Between-item variances and covariances of panel means also use $\mathrm{ddof} = 1$. With unequal ratings per item, the idiosyncratic term σ_p^2/k_p is replaced by $\hat{\sigma}_p^2$ times the mean of $1/k_{p,i}$, and only items with $k_{p,i} \geq 2$ in both pools enter a pair. For DICES-990, coding Unsure as 0 or 1 instead of

Table 9: Real-data analyses: preprocessing, estimator settings and resampling.

dataset	source and preprocessing	panels, coding, estimator	resampling
SummEval	`model_annotations.aligned.jsonl` (1,600 summaries, 100 documents), obtained from the download link in the repository README	experts $k = 3$ vs. crowd $k = 5$, scores 1–5; one-way ANOVA per panel; within-document ceilings on document-centred panel means ($n - D$ degrees of freedom, $D = 100$)	Table 2: 1,000 item draws; ceilings: 1,000 document draws
DICES-990	`diverse_safety_adversarial_dialog_990.csv`, column `Q_overall`	US (80 raters) vs. India (92); No/Unsure/Yes $= 0/0.5/1$; unequal k: $\hat{\sigma}_p^2 \cdot \overline{1/k_{p,i}}$	500 item draws
D3code	`d3-ratings.csv`, `d3-raters.csv`; rating -1 dropped	8 regions, 28 pairs; raw 0–4; items with $k \geq 2$ in both regions; unequal k as above	500 item draws per pair
MultiPref	`allenai/multipref`, default split (10,461 comparisons); fields `overall_pref`, `helpful_pref`, `truthful_pref`, `harmless_pref`	2 experts vs. 2 non-experts (disjoint IDs); A clearly $+2$, slightly $+1$, tie 0, B slightly -1, clearly -2	1,000 item draws
IRep	`IREP_data_release.csv`; International pool excluded	3 city pools, 4,248 items with 2 raters in each; 31 labels (no `error.other`); pool covariance = mean of its two pairs	500 item draws
Jigsaw SRP	Kaggle release (Google Jigsaw, 2022); null ratings dropped; 25,490 comments with $k \geq 2$ in all pools	3 pools; `toxic_score` $-2, -1, 0, 1$ as continuous; unequal k; pool covariance = mean of its two pairs	500 item draws
Cert.–fidelity	all pairs above	sign of $\rho_{k,p} - \rho_{k,q}$ vs. $\rho_p^\star - \rho_q^\star$; D3code also with Bonferroni over 28 pairs	500 paired item draws (4,000 for Bonferroni)
Metric correlations	Table 1 of Liu et al. (2023)	cited, not recomputed (summary-level Spearman, protocol of Zhong et al. (2022))	none
G-Eval vs. both panels	G-Eval release, `results/gpt4_*_detailed.json`	score = mean over the 20 responses of the first number in each (responses without a number skipped); join on (doc_id, system_id) = (id, model_id); per-document Spearman over 16 systems, averaged over the documents where it is defined (constant scores skipped)	1,000 document draws
Nine judges vs. both panels	`lguerdan/indeterminacy-experiments`, 200 summaries per dimension	released response table, indexed by item, forced-choice (FC) option, response set (RS) and draw: positive-option count summed over RS sets and draws, over all recorded answers (10 draws, 8 for o3-mini); items matched by id and summary text (first record in file order for the 7 duplicated texts); modal check positive when more than half of the answers are positive (a tie counts as negative); null answers kept in the denominator; Pearson with each panel mean	1,000 item draws
GPT-4 vs. MultiPref panels	`allenai/multipref`, configuration `gpt4_overall_binarized`	judge $= +1$ if its chosen completion is completion A, -1 if B (text match of the last message of `chosen`, all 10,461 matched); Pearson with each panel mean	1,000 item draws

0.5 gives $r = 0.81$–0.83, both pools flagged, and 92–94% shared variance. Data sources are listed in Table 11; the Jigsaw Specialized Rater Pools are released on Kaggle (Google Jigsaw, 2022).

Table 10: Synthetic results: model parameters and replications ($\mu_i \sim \mathcal{N}(0,1)$, $\sigma_s^2 = 1$ throughout).

result	parameters	replications / output
Inflation curve (Fig. 1a)	$\tau^2 = 0.36$, $\sigma^2 = 1.47$, $k \in \{2, 5, 20, 100\}$	analytic: $0.649, 0.822, 0.949, 0.989$; limit 0.735
Calibration (Fig. 1b)	2 panels, $k = 4$, $n = 200$, $\sigma^2 = 1$, $\tau^2 \in \{0, 0.3\}$ added to both panels; panel 1 tested	500 replications, 250 draws: FPR 4.8%, power 93.4%
Calibration, robustness (§3.2)	as above; ratings rounded to 1–5 ($\mathrm{round}(3 + y)$, clipped), or $k_{p,i}$ drawn per item uniformly in $\{2, \ldots, 6\}$ with that many ratings generated (no rater removed), or both; each setting and each τ^2 with its own `default_rng(0)`	400 replications, 250 draws: FPR 5.5%, 5.5%, 6.8%; power 93.8%, 93.8%, 94.2% (Monte Carlo standard error 1.1–1.3 points)
Prop. 2 check	$n = 300$, $k = 4$, $\sigma^2 = 1$, $\tau^2 \in \{0.1, 0.3, 0.6, 1.0\}$	1,000 replications: $\lvert\mathrm{Cov} - \sigma_s^2\rvert \leq 0.4\%$
Prop. 4 check	3 panels, $n = 300$, $k = 4$, $\tau^2 = 0.3$, $\sigma^2 = 1$; two share $\gamma = 0.3$	1,000 replications: 1.30 vs. 1.00
Variance formula (Prop. 5)	(n, P, k, τ^2): $(300, 2, 4, 0.3)$, $(300, 4, 4, 0.3)$, $(300, 3, 6, 0.6)$; $\sigma^2 = 1$ known	2,000 replications: $0.98, 1.03, 1.06\times$ formula
Per-panel detection optimum (App. F)	$b = 24$, $n = 200$, $\sigma^2 = 1$, $\lambda \in \{0.1, 0.3, 0.5, 1.0\}$, $P \in \{2, 3, 4, 6, 8, 12\}$; SNR $= \tau^2/$s.d. of $\hat{\tau}_p^2$ over replications, σ^2 estimated	1,500 replications: best $P = 2, 3, 3, 3$
DesignPanels check (Alg. 4)	$b = 24$, $n = 200$, $\sigma^2 = 1$, $\lambda \in \{0.1, 0.3, 0.5, 1.0\}$; price at $\tau^2 = 0.3$	analytic: pooled $P = 4, 8, 12, 12$; price $0.9486 - 0.9375$
ItemsNeeded check (Alg. 5)	20 pilots of 200 items, $\tau^2 = 0.3$, $k = 4$, $\gamma = 0.3$, $\alpha = 0.05$, $1 - \beta = 0.8$; 250 draws	median $n = 135$; simulated power at $n = 135$: 84% (200 replications, 200 draws)
CovarianceCheck on Jigsaw (Alg. 6)	three pools, Bonferroni over 3 pairs, $\alpha = 0.10$	250 draws: only LGBTQ–Control elevated (lower bound 0.025)
CovarianceCheck in simulation (Alg. 6)	3 panels, $n = 300$, $k = 4$, $\tau^2 = 0.3$, $\sigma^2 = 1$; panels 1–2 share $\gamma = 0.3$, or none; Bonferroni over 3 pairs, $\alpha = 0.10$	200 replications, 250 draws: sharing pair 94%, independent pairs 0%; any pair when all independent 6.5%
Allocation (Fig. 2, price)	$\tau^2 = 0.3$ (price) or $\lambda\sigma^2$ (SNR), $\sigma^2 = 1$, $n = 200$, $b = 24$	analytic: $R = 0.9486$ vs. 0.9375

Table 11: Data sources.

source	location
SummEval	`https://github.com/Yale-LILY/SummEval`
DICES	`https://github.com/google-research-datasets/dices-dataset`
D3code	`https://github.com/google-research-datasets/D3CODE`
MultiPref	`https://huggingface.co/datasets/allenai/multipref`
IRep	`https://github.com/google-research-datasets/replication-dataset`
WMT20 DA	`https://github.com/WMT-Metrics-task/wmt20-metrics`
WMT MQM	`https://github.com/google/wmt-mqm-human-evaluation`
G-Eval outputs	`https://github.com/nlpyang/geval`
Nine-judge outputs	`https://huggingface.co/datasets/lguerdan/indeterminacy-experiments`

L Comparison with the closest prior work

Table 12 summarises the differences stated in §2. Every entry restates that section; none is a new claim about the cited works.

Table 12: What each line of work estimates and tests, as stated in §2. –: not discussed there.

	xRR (Wong et al., 2021)	anchor–judge model (Sunkavalli, 2026)	alt-test (Calderon et al., 2025)	this work
object audited	agreement between rater pools	error correlation between an anchor and a judge	a judge against one annotator pool, treated as the target	the human reference itself
unit	a pair of pools (symmetric)	judge side	one pool, leave one annotator out	each panel, at the k the benchmark uses
main quantity	normalized κ_x, read as a disattenuated correlation	shared judge-side component	judge versus held-out annotators	$\hat{\tau}_p^2$, $\rho_p^\star$ and the validity ceiling $\sqrt{\rho_p^\star}$
test	confidence intervals and tests left to future work	null-calibrated 5% thresholds, item-bootstrap intervals	Benjamini–Yekutieli false-discovery guarantee	one-sided bootstrap test with stated size and power (§3.2)
ceiling on judge–human correlation	not applied to a machine evaluator	not defined	–	yes
certificate–fidelity inversion	–	–	–	yes (§4)
design rules	–	–	–	yes (Algorithms 4–5)

M Census: which released datasets admit the two-panel test

Frame (reproducible). We consider human-annotation datasets that (i) serve as evaluation references, preference data or subjective-judgement benchmarks in LLM/NLP work, and (ii) release *individual annotator-level* labels, or were examined and found not to release them. They are drawn from the diverse-annotator and learning-with-disagreement literature we cite, accessed September 2026. For each, we read its structure mechanically from the released schema. There are three cases. T2: two or more *independent* panels on the same items with $k \geq 2$, which allows the test. T1: a single multi-annotated pool, which allows only the inflation result. T0: a single annotation per item, which allows neither. A third party can apply the same rule to extend the table.

Table 13: Census by panel structure (frame and criteria above; accessed September 2026). T2 rows admit the calibrated test as released.

dataset	structure	tier
SummEval (Fabbri et al., 2021)	expert + crowd, same items	T2
DICES-990 (Aroyo et al., 2023)	US + India pools, same items	T2
D3code (Mostafazadeh Davani et al., 2024)	8 geo-cultural regions, same items	T2
IRep (Wong et al., 2021)	3 city pools, same items (k=2)	T2
MultiPref (Miranda et al., 2025)	expert + non-expert, disjoint, same items	T2
Jigsaw SRP (Goyal et al., 2022; Google Jigsaw, 2022)	3 pools, 8 shared raters	T2[†]
DICES-350 (Aroyo et al., 2023)	single US pool, fully crossed	T1[‡]
POPQUORN (Pei & Jurgens, 2023), NLPositionality (Santy et al., 2023), ChaosNLI (Nie et al., 2020), Kumar (Kumar et al., 2021), HANNA (Chhun et al., 2022), WMT20 MQM	single multi-annotated pool	T1
WMT20 DA[§]	one score / item	T0

[†] Jigsaw allows the test only under an independence caveat (shared raters across pools). [‡] DICES-350's expert safety label is a single gold value per item (`safety_gold` in the release), not a set of expert ratings, so it cannot serve as a second panel. Its fully crossed crowd pool could in principle be split demographically into sub-panels on the same items. We did not perform that split, so we classify it T1: we count only designs that allow the test *as released*. IRep labels facial-expression videos rather than text; we include it as the dataset released with xRR, the closest prior work. [§] The

released WMT20 segment-level DA files give one mean score per segment with its rating count (one rating for 49% of en–de and 83% of zh–en segments) and no individual ratings, so the within-panel variance cannot be estimated. The table lists every dataset we examined. Of these 14, exactly five are T2 as released, with Jigsaw a sixth under an independence caveat; the other eight give a single pool (T1) or a single score per item (T0). The T2 panels mostly contrast populations (locale, culture, region). A flagged $\hat{\tau}^2$ there is a population difference: a legitimate property of the reference that changes how it should be read, not annotator error.