Agreement Metrics for LLM-as-Judge Evaluation: What to Report and Why

Anonymous ACL submission

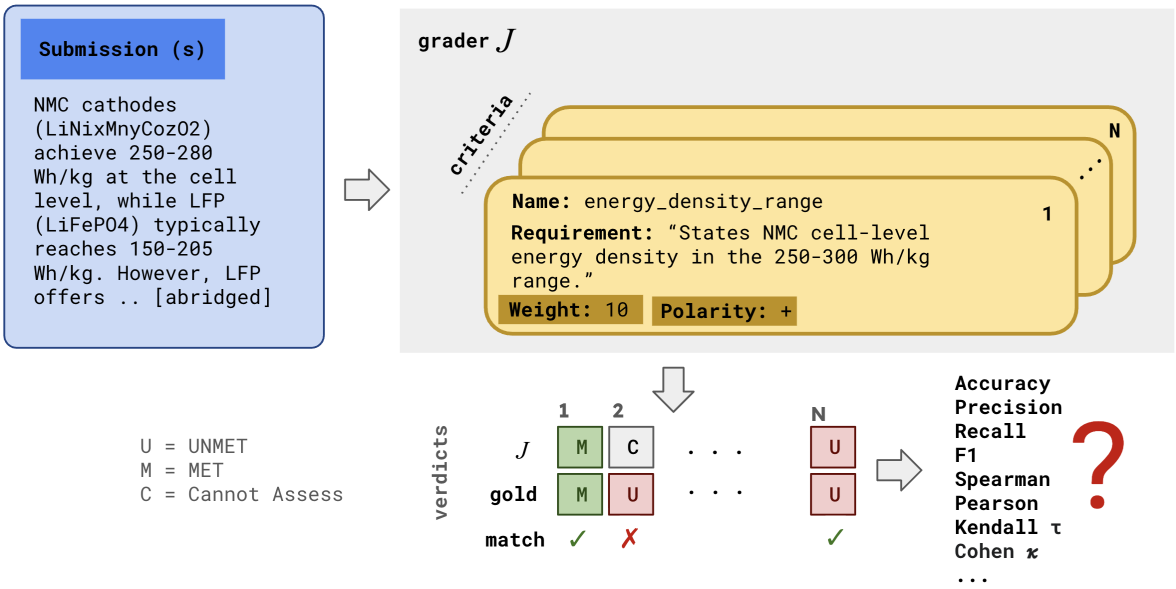

Figure 1: Validating an LLM judge yields many agreement coefficients. This paper asks which are independent and which merely restate one another, how judges that abstain complicate the analysis, and how to handle multiple judges.

Abstract

Validating an LLM judge against human annotations usually means reporting several agreement statistics: accuracy, precision, recall, F_1, Cohen's κ, and one or more rank correlations. A survey of 24 recent LLM-as-judge papers finds metric choice entangled with the judgment scale, tie handling, invalid outputs, and abstention handling, and those choices rarely stated. For binary criteria — the common case in rubric-based evaluation, where each criterion is graded `MET` or `UNMET` — most of the reported numbers are redundant: Pearson's r, Spearman's ρ, Kendall's τ_b, the phi coefficient ϕ, and the Matthews Correlation Coefficient all reduce to a single number on non-degenerate binary data, so reporting several of them only creates an illusion of corroborating evidence. Cohen's κ is the one agreement coefficient that adds information: it shares ϕ's numerator but normalizes differently, and the gap between them measures how far the judge's positive-label rate has drifted from the human's. We then trace what changes when a judge may abstain with a `CANNOT_ASSESS` verdict: the three common ways of handling abstentions are not interchangeable preprocessing choices but answer different questions, and they break the binary equivalences. The same equivalences reappear, up to a negligible finite-sample correction, for multi-judge ensembles scored with Fleiss' κ or Krippendorff's α. We close with a reporting checklist that names the judgment scale, the abstention and tie handling mode, coverage, the confusion matrix, and the aggregation level alongside any scalar agreement coefficient.

1 Introduction

An LLM judge earns trust only after it is validated against human annotations, and that validation reaches the reader as a set of agreement statistics: accuracy, precision, recall, F_1, Cohen's κ, and one or more rank correlations (Pearson, Spearman, Kendall). A common case is *rubric-based evaluation*, where a holistic judgment is decomposed into independently graded criteria, each typically binary (`MET`/`UNMET`) (Hashemi et al., 2024; Saad-Falcon et al., 2023). To see how these statistics are reported in practice, we surveyed 24 recent LLM-as-judge and LLM-evaluator papers (full per-system coding in Appendix A). Metric

choice tracks the judgment scale: several rank correlations are reported together for graded or continuous scores (e.g., Liu et al., 2023; Ye et al., 2024; Kim et al., 2024), while accuracy, agreement, win rate, F_1, or κ are used for binary or pairwise verdicts (e.g., Zheng et al., 2023; Verga et al., 2024; Lee et al., 2026). The handling of ties, invalid outputs, refusals, and insufficient-information (CANNOT_ASSESS) responses[1] is heterogeneous — sometimes excluded, sometimes folded into a valid label — and almost never reported as a deliberate choice.

What current papers do not articulate is what these choices cost. On binary verdicts, are several reported coefficients independent evidence or one number computed several ways? When a judge may abstain, does the resolution of that abstention change the quantity being estimated, or is it harmless preprocessing? The relationships that settle such questions are classical (Cohen, 1960; Matthews, 1975; Warrens, 2008), but they are rarely connected to how LLM judges are scored, and abstention is treated as a preprocessing detail rather than a modeling choice. The gap is not missing mathematics but a missing account of what the reported numbers mean on the scales LLM judges actually produce.

This paper supplies that account. It is an analysis with reporting guidelines, not a comprehensive survey: the contributions are an audit of how 24 recent systems report agreement (Appendix A), a unifying account of which binary agreement statistics are redundant and what their differences measure, an estimand-level treatment of the CANNOT_ASSESS verdict, and a reporting checklist (Section 8). The analysis needs nothing beyond the 2×2 confusion matrix a validation already produces. First, on non-degenerate binary verdicts, Pearson's r, Spearman's ρ, Kendall's τ_b, ϕ, and MCC are a single statistic, and Cohen's κ differs from it only through a marginal-sensitive normalization; we draw out the reporting consequences (Sections 3–5). Second, the CANNOT_ASSESS verdict is a choice of estimand, not a preprocessing step: the three handling modes answer different questions — a selective-prediction view (Chow, 1970; El-Yaniv and Wiener, 2010; Kamath et al., 2020) — and the binary equivalence breaks once a third class is admitted (Section 6). Third, the equivalence extends to multi-judge ensembles, where Fleiss' κ, Krippendorff's α, and average pairwise ϕ agree up to a negligible finite-sample correction (Section 7). We close with a reporting checklist (Section 8) that names the judgment scale, the abstention and tie handling mode, coverage, the confusion matrix, and the aggregation level alongside any scalar.

[1] Equivalent labels in the literature include IDK, REFUSE, and ABSTAIN.

2 Preliminaries

Consider the binary classification problem induced by per-criterion judgments. Let $y_i \in \{0, 1\}$ denote the ground-truth label for item $i \in \{1, \ldots, N\}$, with 1 encoding the positive verdict (e.g., MET) and 0 the negative verdict (UNMET); let $\hat{y}_i$ denote the LLM judge's prediction. Abstention is deferred to Section 6. Each cell of the 2×2 confusion matrix is a dataset-level count: the number of items i whose ground-truth label y_i and predicted label $\hat{y}_i$ equal the row and column values. Writing this count as $n_{ab} = |\{i : y_i = a,\ \hat{y}_i = b\}|$, the four cells are $TP = n_{11}$, $FN = n_{10}$, $FP = n_{01}$, and $TN = n_{00}$:

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	TP	FN
$y = 0$	FP	TN

with $N = TP + FN + FP + TN$ the total number of items. Denote $\pi = (TP + FN)/N$ for the true prevalence and $\hat{\pi} = (TP + FP)/N$ for the predicted positive rate.

2.1 Judgment Scale and Scope

The identities below are exact for binary vectors; they do not make Pearson, Spearman, and Kendall redundant on graded, ordinal, or continuous scores. The first reporting decision is therefore what judgment scale the evaluator has produced, not which metric to compute. Table 1 names the five scales encountered in recent work and the metric framing each calls for.

The remainder of the paper develops the binary row (Sections 3–5) and, once abstention is admitted, the nominal row (Section 6); the ordinal and continuous scales lie outside the scope of these results.

3 Closed-Form Metric Definitions

The standard definitions, expressed directly in terms of TP, FP, FN, TN:

Accuracy:

$$\text{Acc} = \frac{TP + TN}{N} \equiv p_o. \tag{1}$$

Judgment scale	Examples	Recommended metric framing
Binary verdict	MET/UNMET criterion decisions	Accuracy, F_1, κ, and one — not several — of ϕ/MCC/Pearson/Spearman/Kendall's τ_b
Pairwise preference	A wins / B wins / tie	State the tie-handling convention explicitly; once reduced to a binary outcome, the equivalences of Section 4 apply
Ordinal score	1–5 rubric ratings	Rank correlations no longer collapse; prefer weighted κ or another ordinal-aware agreement coefficient
Continuous/direct assessment	Scalar quality scores	Pearson, Spearman, and Kendall answer different questions and may all be informative
Nominal with abstention	MET/UNMET/CA	Use a three-class agreement statistic with an explicit weight matrix, or state the reduction mode used to collapse to two

Table 1: A judgment-scale taxonomy for LLM-as-judge evaluation. The equivalence results concern the binary row; the nominal-with-abstention row is taken up in Section 6.

F1, precision, and recall (positive class): With precision $P = TP/(TP + FP)$ and recall $R = TP/(TP + FN)$, the F_1 score is their harmonic mean

$$F_1 = \frac{2\,P\,R}{P + R} = \frac{2\,TP}{2\,TP + FP + FN}. \tag{2}$$

F_1 does not depend on TN and is asymmetric under label flipping (MET $\leftrightarrow$ UNMET).[2] Precision and recall are the exception to the collapse below: every symmetric statistic that follows (accuracy, κ, ϕ, the rank correlations, and F_1) is invariant under $FP \leftrightarrow FN$, so none distinguishes a judge that misses positives from one that over-predicts them.

Cohen's κ: With expected agreement $p_e = \pi\hat{\pi} + (1 - \pi)(1 - \hat{\pi})$ (Cohen, 1960),

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \tag{3}$$

Multiplying through by N^2 and canceling shared terms yields

$$\kappa = \frac{2\,(TP \cdot TN - FP \cdot FN)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}. \tag{4}$$

Phi coefficient and MCC: For binary vectors, the Pearson correlation reduces to the phi coefficient (Matthews, 1975),

$$\phi = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)}} \times \frac{1}{\sqrt{(TN + FP)(TN + FN)}}, \tag{5}$$

which is identical to the binary Matthews Correlation Coefficient.

[2]The negative-class analog is $F_1^{(0)} = 2\,TN/(2\,TN + FP + FN)$.

4 Equivalence of Five Metrics on Binary Data

Observation 1

For any pair of binary vectors $(y, \hat{y}) \in \{0, 1\}^N$ with non-degenerate marginals,

$$\rho_{\text{Pearson}}(y, \hat{y}) = \rho_{\text{Spearman}}(y, \hat{y}) = \tau_b(y, \hat{y}) = \phi(y, \hat{y}) = \text{MCC}(y, \hat{y}). \tag{6}$$

The result is classical and follows from standard properties of the underlying statistics, but its reporting consequences (Section 8) are easy to miss in LLM-as-judge work. The Pearson-MCC-phi identity is immediate from Eq. (5). The two non-obvious pieces are why Spearman and Kendall coincide with Pearson on binary data.

Why Spearman equals Pearson on binary data? Spearman's ρ is Pearson's r computed on the rank-transformed inputs. With only two distinct values present, all zeros receive the common average rank $a_0 = (n_0 + 1)/2$ and all ones receive $a_1 = n_0 + (n_1 + 1)/2$, where n_0, n_1 are the counts of zeros and ones. The map $x \mapsto a_0 + (a_1 - a_0)\,x$ is affine, and Pearson's r is invariant under affine rescaling of either input — so $\rho_{\text{Spearman}} = \rho_{\text{Pearson}}$.

Why Kendall's τ_b equals ϕ on binary data? The tie-corrected form (Kendall, 1945) is

$$\tau_b = \frac{C - D}{\sqrt{(n_{\text{tot}} - n_y)(n_{\text{tot}} - n_{\hat{y}})}}, \tag{7}$$

where $n_{\text{tot}} = \binom{N}{2}$ and $n_y, n_{\hat{y}}$ are the numbers of pairs tied on y and on $\hat{y}$ respectively. Reading pair types off the cells of the 2×2 table of Section 2, a concordant pair joins a $(1, 1)$ cell with a $(0, 0)$ cell, so $C = TP \cdot TN$, and a discordant pair joins a $(1, 0)$ cell with a $(0, 1)$ cell, so $D = FN \cdot FP$.

The pairs untied on y number $n_{\text{tot}} - n_y = (TP + FN)(FP + TN)$, and those untied on $\hat{y}$ number $n_{\text{tot}} - n_{\hat{y}} = (TP+FP)(FN+TN)$. Substituting,

$$\tau_b = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(FP+TN)}} \times \frac{1}{\sqrt{(TP+FP)(FN+TN)}} = \phi, \tag{8}$$

matching Eq. (5).

5 Cohen's κ and Marginal Sensitivity

Cohen's κ is close to the common value of Section 4 but not equal to it. The gap is fully controlled by how well the judge's positive-prediction rate matches the human's.

Comparing Eq. (4) and Eq. (5), the numerator $TP \cdot TN - FP \cdot FN$ is shared (up to a factor of 2). The denominators differ:

$$D_\kappa = (TP+FN)(FN+TN) + (TP+FP)(FP+TN), \tag{9}$$

$$D_\phi = 2\sqrt{(TP+FN)(FN+TN)} \times \sqrt{(TP+FP)(FP+TN)}. \tag{10}$$

Writing $A = (TP+FN)(FN+TN)$ and $B = (TP+FP)(FP+TN)$, we have $D_\kappa = A+B$ and $D_\phi = 2\sqrt{AB}$. The AM-GM inequality $A + B \geq 2\sqrt{AB}$, with equality iff $A = B$, gives

$$\frac{\kappa}{\phi} = \frac{2\sqrt{AB}}{A+B} \leq 1, \tag{11}$$

and

$$|\kappa| \leq |\phi|, \tag{12}$$

with equality iff $A = B$, which is equivalent to $\pi = \hat{\pi}$. **Cohen's κ thus penalizes the judge for marginal-distribution mismatch with the ground truth in a way that ϕ does not.** A judge that perfectly orders items but commits to a different positive rate than the human annotator receives a strictly lower κ than ϕ.

A second reformulation rewrites κ as a rescaled accuracy: $\kappa = (\text{Acc} - p_e)/(1 - p_e)$. In the balanced case $\pi = \hat{\pi} = 0.5$, $p_e = 0.5$ and $\kappa = 2\,\text{Acc} - 1$. In the extreme-prevalence case $\pi \to 1$ (or $\pi \to 0$), $p_e \to 1$ and small accuracy gaps drive κ to zero or negative values — the *kappa paradox* discussed by Feinstein and Cicchetti (1990).

Worked example

With $N = 100$ and $(TP, FN, FP, TN) = (40, 10, 20, 30)$, $\pi = 0.50$ and $\hat{\pi} = 0.60$. Direct computation gives Acc $= 0.700$, $F_1 = 0.727$, $p_e = 0.500$, $\kappa = 0.400$, and $\phi = 1000/\sqrt{6 \times 10^6} \approx 0.408$. The $|\kappa| < |\phi|$ gap is small because the marginal mismatch is small. In a low-prevalence regime $(TP, FN, FP, TN) = (5, 5, 0, 90)$ ($\pi = 0.10$), Acc $= 0.95$ but $\kappa \approx 0.643$ and $\phi \approx 0.688$; a degenerate always-negative predictor on the same true distribution would achieve Acc $= 0.90$ with $\kappa = 0$ and undefined ϕ — illustrating the danger of reporting accuracy alone on imbalanced criteria.

Sampling-distribution behavior. The pointwise collapse extends to the first-order sampling behavior of $\hat{\kappa}$ and $\hat{\phi}$ under multinomial sampling of the 2×2 cell counts. When $\pi = \hat{\pi}$ — the matched-marginals condition that makes $\kappa = \phi$ pointwise — the asymptotic variances coincide to leading order:

$$\text{Var}(\hat{\kappa}) = \text{Var}(\hat{\phi}) + O(N^{-2}). \tag{13}$$

The mechanism is a gradient-coincidence argument on the simplex, not a cell-by-cell partial-derivative identity, and the variance gap off matched marginals is second-order in the mismatch; Appendix C gives the argument and a numerical check.

For finite-sample uncertainty bands, bootstrap resampling is preferable to either asymptotic formula, with one caveat. In rubric evaluation the unit of analysis is rarely an isolated criterion decision: criteria are clustered within items (multiple criteria per submission), within prompts or rubrics (multiple items sharing a rubric), and within judges (in an ensemble, multiple decisions by the same model). A flat decision-level resample understates uncertainty when within-cluster correlations are positive, as is the norm when a single judge mistake propagates across the criteria of one item. An item-level or hierarchical cluster bootstrap (resample items, then optionally resample criteria within items) is preferable; standard cluster-bootstrap implementations are available in the survey-statistics literature (e.g., Field and Welsh, 2007).

5.1 Micro versus Macro Aggregation across Rubric Criteria

Rubric evaluation is not a flat sequence of independent decisions; the same hierarchy that underlies the cluster-bootstrap caveat forces a second choice, about how to aggregate the per-decision metric into a single number. Three levels are standard. *Micro-averaging* pools all item $\times$ criterion decisions into

one long sequence and computes the metric — accuracy, κ, F_1 — on the pooled 2×2 table. *Macro-averaging* computes the metric separately for each criterion and averages the per-criterion numbers, unweighted. *Item-level aggregation* first reduces each item's vector of criterion verdicts to a single per-item verdict (a weighted-sum score thresholded into pass/fail, or an all-criteria-met conjunction, depending on the rubric) and computes agreement on the resulting per-item labels.

These numbers estimate different quantities, and a scalar agreement figure reported without naming its aggregation level is ambiguous. Consider a two-criterion rubric in which criterion A fires on 80% of items and is easy ($\kappa = 0.9$) and criterion B fires on 5% of items and is hard ($\kappa = 0.2$). Micro-κ is dominated by A's decisions, landing near 0.85; macro-κ averages to 0.55; item-level κ depends on the rubric's weights and threshold, and may track either criterion or neither. This choice interacts with the resampling unit above: a criterion-level cluster bootstrap matches macro-averaging, an item-level cluster bootstrap matches item-level aggregation, and a flat decision-level bootstrap matches only micro-averaging, and then only under the (usually false) independence assumption. Our recommendation is to report all three when the rubric has more than a handful of criteria, and name the chosen level whenever a single number is reported.

6 The Abstention Complication

The preceding analysis assumed every prediction and every ground-truth label was strictly binary. LLM judges, however, are commonly prompted to emit a third verdict when the submission does not contain sufficient information to assess the criterion. We refer to this verdict as `CANNOT_ASSESS` (`CA`).[3] Agreement metrics are computed only after these verdicts are resolved.

6.1 Three Modes, Three Estimands

Let $\mathcal{D} = \{(y_i, \hat{y}_i)\}_{i=1}^{N}$ with $y_i, \hat{y}_i \in \{\texttt{MET}, \texttt{UNMET}, \texttt{CA}\}$. Three resolutions are in common use — *exclude* the abstentions, *recode* them as negative, or keep them as a *third class*. The three paragraphs below treat each in turn; the point is that they are not interchangeable preprocessing steps but estimate different population quantities.

[3]Human annotators produce `CANNOT_ASSESS` labels as well, though the underlying causes differ: a judge typically abstains for lack of evidence in the submission, an annotator when the criterion itself is ambiguous for the item.

Exclude: To make the abstention mechanism explicit, introduce latent binary verdicts $y_i^*, \hat{y}_i^* \in \{0, 1\}$ — the decisions the human and judge would have committed to if neither were permitted to abstain — together with abstention indicators $m_i, \hat{m}_i \in \{0, 1\}$, where $m_i = 1$ iff the human abstained (and analogously for $\hat{m}_i$). The observed verdicts are $y_i = y_i^*$ when $m_i = 0$ and $y_i = \texttt{CA}$ when $m_i = 1$, with the symmetric definition for $\hat{y}_i$. The exclude mode discards every pair with $m_i = 1$ or $\hat{m}_i = 1$ and computes the binary metric on the residue; writing T for any of accuracy, F_1, κ, ϕ viewed as a functional of the joint latent distribution, the estimand is

$$T\big(\mathbb{P}_{(y^*, \hat{y}^*) \,|\, m=0 \,\wedge\, \hat{m}=0}\big). \tag{14}$$

This equals the unconditional latent functional $T(\mathbb{P}_{(y^*, \hat{y}^*)})$ only under assumptions about the joint abstention mechanism — typically missing-at-random conditional on observed item features, or the stronger missing-completely-at-random assumption (Little and Rubin, 2019). The exclude estimand is the *selective-accuracy* measure of the binary judge, restricted to the $(1 - \Pr[m = 1])(1 - \Pr[\hat{m} = 1])$ fraction of the population on which both sides chose to commit; it characterizes performance on the covered subset, not on the full deployment distribution.

Recode-as-negative: Map `CANNOT_ASSESS` to `UNMET` in both vectors, then compute the binary metric. Under a plain misclassification loss the `CA` and `UNMET` responses become indistinguishable, but the calibration information the `CA` verdict was introduced to carry is destroyed at the measurement stage. The sign of the resulting bias depends on whether the criterion is positively or negatively weighted.

Three-class: Retain `CANNOT_ASSESS` as a distinct category and report a 3×3 agreement metric such as multi-class accuracy, multi-class Cohen's κ (Conger, 1980), or a weighted variant with an explicit distance matrix encoding the geometry of disagreement.

Appendix B works a single 3×3 table through all three modes, showing that they yield different accuracy, κ, and F_1 values from one set of judgments.

6.2 The Equivalence Breaks at $k = 3$

The affine-rank argument behind Observation 1 required the input vectors to take exactly two distinct

values. With three values $\{0, \frac{1}{2}, 1\}$ (the ordinal encoding with CA between MET and UNMET), the rank map sends $0 \mapsto (n_0+1)/2$, $\frac{1}{2} \mapsto n_0+(n_{1/2}+1)/2$, and $1 \mapsto n_0 + n_{1/2} + (n_1 + 1)/2$. Direct calculation shows this map is affine in the input value iff $n_0 = n_1$ — iff MET and UNMET prevalences coincide. Generically, $\rho_{\text{Spearman}} \neq \rho_{\text{Pearson}}$ on three-class data. The phi coefficient has no canonical 3×3 analog; the multi-class MCC extension of Gorodkin (2004) does not reduce to either Pearson on a chosen encoding or to Kendall's τ_b. The choice among multi-class agreement statistics is therefore a substantive modeling decision.[4]

6.3 Selective Prediction and the Abstention Rate as a Metric

Under recode-as-negative, a judge that abstains on uncertain negatives and one that commits to UNMET on the same items receive identical scores. Whether this *incentivizes* one behavior over the other depends on the judge's training objective, outside the scope of the metric itself — but under an objective that explicitly rewards calibrated abstention (as in selective prediction), the recoding erases the very signal the objective conditions on. The selective-prediction literature (Chow, 1970; El-Yaniv and Wiener, 2010) treats this by reporting a two-axis summary — *coverage* (the fraction of non-abstention predictions) and *selective accuracy* (the accuracy on the covered subset) — rather than a single scalar; recent NLP work develops the same view for question answering (Kamath et al., 2020), classification (Xin et al., 2021), and calibrated LLM uncertainty (Kadavath et al., 2022).

7 Multi-Judge Ensembles

Ensembles of judges reduce the prompt-sensitivity and model-specific biases of a single judge (Wang et al., 2023a; Verga et al., 2024; Bavaresco et al., 2025), and they add a measurement object single-judge setups lack: *inter-judge agreement*, which flags items the rubric does not unambiguously resolve, independently of agreement with ground truth. With $R \geq 2$ judges grading each criterion (Rao and Callison-Burch, 2026), the standard inter-judge statistics are Fleiss' κ (Fleiss, 1971) and Krippendorff's α (Krippendorff, 2004).

[4]Observation 1 also fails on binary data with degenerate marginals; the $k = 3$ break is qualitatively different in that it cannot be repaired by any choice of encoding.

7.1 Binary Ensembles

Let $\hat{y}_{ij} \in \{0, 1\}$ denote judge j's verdict on item i, and let $r_i = \sum_{j=1}^{R} \hat{y}_{ij}$. With global rates $p_1 = (NR)^{-1} \sum_i r_i$ and $p_0 = 1 - p_1$, Fleiss' κ has observed pairwise agreement

$$P_o = \frac{1}{NR(R-1)} \sum_{i=1}^{N} \big[r_i(r_i - 1) + (R - r_i)(R - r_i - 1) \big] \tag{15}$$

and expected agreement $P_e = p_0^2 + p_1^2$, giving $\kappa_F = (P_o - P_e)/(1 - P_e)$. Krippendorff's α on the same data uses observed disagreement $D_o = 1 - P_o$ and expected disagreement $D_e = (1 - P_e) \cdot NR/(NR-1)$, incorporating a finite-sample correction.

These statistics give the binary multi-judge analog of Observation 1.[5]

Observation 2

On binary data with R judges and N items, the inter-rater agreement statistics satisfy the exact identity

$$\alpha = \kappa_F + \frac{1 - \kappa_F}{NR}, \tag{16}$$

so $\alpha \geq \kappa_F$ with equality iff $\kappa_F = 1$, and $\alpha \to \kappa_F$ as $N \to \infty$. The average pairwise Pearson, Spearman, Kendall's τ_b, ϕ, and MCC over the $\binom{R}{2}$ judge pairs are all equal, and coincide with κ_F when all judges share the same marginal MET rate.

For typical values ($\kappa_F = 0.6$, $N = 100$, $R = 3$), $\alpha - \kappa_F \approx 1.3 \times 10^{-3}$, below the precision at which agreement coefficients are reported. A binary ensemble-validation report should pick one of $\{\kappa_F, \alpha, \text{average pairwise } \phi\}$; the others add no information beyond a finite-sample correction.

7.2 Three Classes: Admitting Abstention

Admitting CANNOT_ASSESS breaks the equivalences of Observation 2 in three ways, each by the mechanism of Section 6.2.

[5]Substituting $D_o = 1 - P_o$ and $D_e = (1 - P_e)\, NR/(NR - 1)$ into $\alpha = 1 - D_o/D_e$ yields $\alpha = 1 - (NR - 1)(1 - \kappa_F)/NR = \kappa_F + (1 - \kappa_F)/(NR)$. The pairwise correlation equalities follow from Observation 1 applied to each judge pair (equality on each pair propagates through any symmetric aggregation), and the equality with κ_F under matched marginals is the binary specialization of the kappa-decomposition of Conger (1980).

Average pairwise Pearson requires a numeric encoding of {MET, UNMET, CA}; the result depends on the encoding (e.g., $\{1, 0, 1/2\}$ versus deleting CA pairs), and no encoding is canonical.

Fleiss' κ on three nominal categories penalizes MET–UNMET and MET–CA confusions equally, while Krippendorff's α with a quadratic-ordinal δ-function penalizes MET–UNMET more than MET–CA, encoding the assumption that CA is "between." On the same ensemble, the two statistics can rank judges differently.

Average pairwise Spearman and τ_b on three-valued data no longer coincide; their divergence is governed by the per-item distribution of CA votes across judges.

Operationally, the multi-judge three-class case admits no single defensible agreement scalar without an explicit weight matrix $W \in \mathbb{R}^{3\times 3}$ specifying the cost of each disagreement. Report per-judge 3×3 confusion matrices alongside any scalar summary, allowing readers to recompute weighted κ under their preferred W.

8 A Checklist for Reporting Agreement Metrics

The audit (Appendix A) found that agreement statistics are frequently reported without the information needed to interpret them, and Sections 3–7 explain why that information is not optional: on binary data the same scalar can stand for several distinct statistics, and its value shifts with the judgment scale, the handling of ties and abstentions, and the aggregation level. The eleven practices below make those choices explicit, so that a reported number is reproducible and comparable across studies. Each follows directly from a result above and costs nothing beyond the confusion-matrix counts a validation already produces; the first is a precondition for the rest.

1. **Name the judgment scale before naming the metric.** State which of the five scales of Table 1 the judge produces: binary verdicts, pairwise preferences (with or without ties), ordinal scores, continuous/direct-assessment scores, or nominal labels with CANNOT_ASSESS / tie / invalid outcomes. Almost every other entry depends on this declaration, because the Section 4 equivalences are exact only after the variables are binary.
2. **On binary data, report at most one of Pearson, Spearman, Kendall's τ_b, ϕ, and MCC.** By Observation 1 these five statistics are identical on non-degenerate binary vectors; prefer ϕ/MCC when the goal is binary association, since the name itself signals the scale. Reporting more than one creates a misleading appearance of triangulated evidence.
3. **Report accuracy together with a marginal-sensitive agreement measure.** Accuracy alone is unreliable on imbalanced criteria, as the kappa-paradox example of Section 5 shows; Cohen's κ is the companion of choice because its gap from ϕ/MCC reflects the mismatch between the judge's and the human's positive-label rates.
4. **Report the 2×2 confusion matrix, or at minimum precision and recall.** As noted in Section 3, every symmetric statistic in this paper — accuracy, F_1, κ, ϕ/MCC, and the rank correlations — is invariant under $FP \leftrightarrow FN$, so none distinguishes a judge that misses positives from one that over-predicts them. For LLM-as-judge validation, judge strictness or leniency is often the most actionable diagnosis.
5. **State how degenerate criteria are handled.** Rubric criteria are often rare, and a criterion with constant human labels or constant judge predictions leaves Pearson, ϕ/MCC, Spearman, and Kendall undefined (and κ uninformative). Such criteria should be reported with their counts and marked NA, rather than silently dropped or coerced to zero.
6. **Name the CA / tie / invalid handling mode in every reported number.** As Section 6 shows, the three modes estimate different quantities; a metric reported without its handling mode is ambiguous among them.
7. **Report abstention, tie, and invalid rates as metrics in their own right.** Following Section 6, a κ of 0.7 on a 40%-abstention sample and a κ of 0.7 on a 5%-abstention sample are not comparable, and abstention drift is often the first observable symptom of prompt instability or distribution shift.
8. **For the exclude mode, report covered-subset performance and coverage together.** The exclusion estimand of Section 6 collapses to the population estimand only under a missing-completely-at-random assumption on abstention; where possible, examine whether abstention correlates with item difficulty, criterion type, prompt, or judge model before

treating the residue as representative.

9. **For three-class or multi-class reporting, specify the weight or cost matrix.** As the three-class worked example in Appendix B shows, scalar summaries over a 3×3 confusion matrix depend on whether CA is treated as closer to MET, closer to UNMET, or equally different from both; without an explicit weight matrix the choice is hidden in the metric's defaults.
10. **Specify the aggregation level.** Rubric evaluation is nested — criteria within items, items within prompts, possibly repeated judgments within models or ensembles — and the Section 5.1 aggregation options (micro, macro, item-level) answer different questions when criteria differ in prevalence or difficulty. State the level explicitly; this is complementary to the cluster-bootstrap caveat in Section 5.
11. **For judge ensembles, report one primary inter-judge agreement statistic and the per-judge confusion matrices.** By Observation 2, $\{\kappa_F, \alpha, \text{average pairwise } \phi\}$ coincide on binary data up to a finite-sample correction of relative order $O(1/(NR))$, so pick one; the per-judge confusion matrices preserve the FP/FN asymmetry of item 4 across the ensemble and become essential when CANNOT_ASSESS is admitted and any scalar summary depends on a non-canonical weight matrix.

9 Conclusions

The variety of agreement statistics in LLM-as-judge validation is largely apparent. Once the judgment scale is fixed to binary verdicts, the six numbers most often reported collapse to four: accuracy, F_1, a single association value ϕ (equal to Pearson, Spearman, Kendall's τ_b, and MCC), and Cohen's κ (Observation 1). Cohen's κ adds information not as a fifth coefficient but through its chance-correction: the κ–ϕ gap measures how far the judge's positive rate has drifted from the human's, closing in value and, to leading order, in sampling variance exactly when the two rates coincide. Reporting Pearson, Spearman, and Kendall side by side then manufactures the appearance of triangulation from one statistic computed three ways.

What carries information beyond this common value is structural rather than numerical. The 2×2 confusion matrix separates a strict judge from a lenient one, a distinction every symmetric coefficient erases. Abstention handling is not preprocessing but a choice of estimand — excluding, recoding, and retaining CANNOT_ASSESS each estimate a different quantity, and the last breaks the binary equivalences. Aggregation level and the ensemble statistic are choices of the same kind: each changes what a number means, not how precisely it is estimated.

An agreement number is therefore interpretable only alongside the declarations that fix its meaning: the judgment scale, the abstention and tie handling mode, the coverage, the confusion matrix, and the aggregation level. None costs more than the counts a validation already computes, and together they let readers reconstruct the equivalent statistics and compare across studies. The checklist of Section 8 collects these declarations; their common lesson is that the first reporting decision is not which coefficient to compute but what was being measured.

Limitations

A few boundaries are worth naming. The analysis treats the observed confusion-matrix counts as fixed; the identities and inequalities hold exactly on any dataset on which they are computed. Whether these counts additionally admit a sampling-distribution interpretation depends on the inferential target: when the evaluation set is itself the object of measurement and the judge is run at temperature zero the counts are deterministic, while generalizing to a wider item population or accounting for judge stochasticity at non-zero temperature places them in the asymptotic regime of Section 5 (or, for finite samples, the cluster bootstrap noted there). The Section 7 multi-judge treatment covers the standard pairwise-aggregable agreement statistics on the unweighted scale; we do not characterize the finite-sample behavior of weighted-κ ensembles under arbitrary weight matrices, nor the behavior of judge ensembles whose members are not exchangeable (for example, one anchored on a frontier model and the rest on small open-weight models). The checklist assumes human ground truth is itself reliable, which is tenuous for the subjective, open-ended tasks where LLM-as-judge methods are most heavily used; analyses of human-judge disagreement (Plank, 2022) suggest some criterion-level disagreement is signal rather than noise.

Generative AI Use Disclosure

The authors acknowledge the use of LLMs in the preparation of this manuscript. Specifically, the authors used Gemini-Pro 3.1 and Claude Sonnet 4.6 to organize the references, proofread the drafts, make plots, and provide light rewrites. The corresponding author has taken care to review the citations and rewrites to ensure they are hallucination free and relevant.

Ethical Considerations

This work analyzes the relationships among agreement metrics and examines how they are reported in literature; it introduces no new data, models, or human-subjects experiments. We see no major ethical considerations arising from the work itself. We note only the indirect point already made in the text: agreement statistics are sometimes reported in ways that overstate how well an automated judge matches human judgment, and the reporting practices recommended here are intended to make such comparisons more transparent.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint*. ArXiv:2306.04757.

C. K. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

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

Anthony J. Conger. 1980. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322–328.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint*.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated evaluation of retrieval augmented generation. *arXiv preprint*. ArXiv:2309.15217.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

C. A. Field and A. H. Welsh. 2007. Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3):369–390.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. 2024. JuStRank: Benchmarking LLM judges for system ranking. *arXiv preprint*. ArXiv:2412.09569; to appear ACL 2025.

J. Gorodkin. 2004. Comparing two k-category assignments by a k-category correlation coefficient. *Computational Biology and Chemistry*, 28(5-6):367–374.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834. Association for Computational Linguistics.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. TIGERScore: Towards building explainable metric for all text generation tasks. *arXiv preprint*. ArXiv:2310.00752; published in TMLR 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and

17 others. 2022. Language models (Mostly) know what they know. *arXiv preprint*. ArXiv:2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint*. ArXiv:2310.08491.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. SAGE Publications, Thousand Oaks, CA.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy yong Sohn, and Kangwook Lee. 2026. How to correctly report llm-as-a-judge evaluations. *Preprint*, arXiv:2511.21140.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint*. ArXiv:2310.05470.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint*. ArXiv:2305.13711.

Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*, 3rd edition. John Wiley & Sons, Hoboken, NJ.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint*.

B. W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Delip Rao and Chris Callison-Burch. 2026. Autorubric: Unifying rubric-based llm evaluation. *Preprint*, arXiv:2603.00077.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. ARES: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint*. ArXiv:2311.09476.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023b. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. *arXiv preprint*. ArXiv:2306.05087.

Matthijs J. Warrens. 2008. On association coefficients for 2x2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73(4):777–789.

Meng-Chen Wu, Md Mosharaf Hossain, Tess Wood, Shayan Ali Akbar, Si-Chi Chin, and Erwin Cornejo. 2025. SEEval: Advancing LLM text evaluation efficiency and accuracy through self-explanation prompting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7372–7383, Albuquerque, New Mexico. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. 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), pages 1040–1051, Online. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. INSTRUCTSCORE: Explainable text generation evaluation with fine-grained feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. ArXiv:2305.14282.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint*. ArXiv:2310.07641.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pages 46595–46623.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. JudgeLM: Fine-tuned large language models are scalable judges. *arXiv preprint*. ArXiv:2310.17631.

A Audited Systems and Reporting Practices

Table A1 gives the full audit summarized in Section 1: the 24 surveyed LLM-as-judge and LLM-evaluator systems, coded by judged output scale, reported metrics, and primary relevance category.

B Worked Example: The Three Modes Compared

Consider the 3×3 confusion table below, with row totals as true marginals and column totals as predicted marginals.

$y \backslash \hat{y}$	MET	UNMET	CA	Total
MET	30	10	5	45
UNMET	10	20	5	35
CA	5	5	10	20
Total	45	35	20	100

Applying each abstention-handling mode to these data:

Mode	N_{eff}	Acc	κ	$F_1^{(\text{MET})}$
Exclude	70	0.714	0.416	0.750
Recode→UNMET	100	0.700	0.394	0.667
Three-class	100	0.600	0.370	n/a

The binary $F_1^{(\text{MET})}$ column applies only after a binary reduction; the three-class row reports no binary reduction. The three modes produce different answers from one set of judgments. They estimate different quantities, corresponding to different questions about the judge. The "n/a" entry in the three-class row reflects that no binary reduction is applied; a multiclass setting still admits per-class F_1, macro-F_1, and micro-F_1 aggregations, none of which is the same statistic as the binary F_1 in the first two rows. The one-vs-rest MET-class F_1 here is $2 \cdot 30/(2 \cdot 30 + 15 + 15) = 60/90 \approx 0.667$, which numerically matches the recode-as-UNMET row by coincidence on this table — the two are different statistics whose values happen to agree here.

C Asymptotic Variance of $\hat{\kappa}$ and $\hat{\phi}$

This appendix supports the matched-marginals variance identity of Eq. (13). On the matched-marginals slice $\{p_{10} = p_{01}\}$, both metrics collapse to the same function of cell probabilities, $\kappa = \phi = 1 - p_{10}/(p_{1\cdot}p_{0\cdot})$, so $\nabla\kappa$ and $\nabla\phi$ agree along the two tangent directions of that slice. The slice-orthogonal direction $(0, +1, -1, 0)$ is the $p_{10} \leftrightarrow p_{01}$ swap, under which both metrics are symmetric and stationary, so the two four-dimensional gradients can differ only along the simplex normal $\mathbf{1}$. Since the multinomial covariance $\Sigma = N^{-1}(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)$ satisfies $\Sigma\mathbf{1} = 0$, the delta-method quadratic forms $\nabla\kappa^\top\Sigma\nabla\kappa$ and $\nabla\phi^\top\Sigma\nabla\phi$ agree to leading order in $1/N$.

As a numerical check, at the matched point $(0.3, 0.1, 0.1, 0.5)$ — where $\pi = \hat{\pi} = 0.4$ — the cell probabilities have unequal row and column sums while the *marginals* π and $\hat{\pi}$ remain equal; direct computation gives $\nabla\kappa - \nabla\phi = (5/6)\,\mathbf{1}$, confirming the all-ones structure away from the $\pi = 0.5$ balanced case. At the balanced matched point $(p_{11}, p_{10}, p_{01}, p_{00}) = (0.4, 0.1, 0.1, 0.4)$ — which adds $\pi = 0.5$ on top of matched marginals — $\kappa = \phi = 0.6$, and the multinomial delta method using the closed-form $\hat{\kappa}$ variance of Fleiss et al. (1969) gives $\text{Var}(\hat{\kappa}) = \text{Var}(\hat{\phi}) = 16/(25N) = 0.640/N$ — not the $0.41/N$ that the bivariate-normal approximation $(1 - \phi^2)^2/N$ would yield for continuous Pearson, which does not apply to Bernoulli cells. Off matched marginals the two variances diverge by a second-order term:

#	Paper / system	Judged output scale	Metrics / reporting practice	Relevance to this paper
1	G-Eval (Liu et al., 2023)	Graded NLG scores	Reports rank/correlation alignment with human judgments.	GRADED Multiple correlations are on graded scores, not raw binary verdicts.
2	LLM-Eval (Lin and Chen, 2023)	Numeric multi-dimensional conversation scores	Reports Pearson/Spearman-style score correlations.	GRADED Not a binary setting.
3	SEEval (Wu et al., 2025)	Scalar NLG quality scores	Reports Pearson/Spearman or Spearman/Kendall depending on benchmark.	GRADED Not a binary setting.
4	CheckEval (Lee et al., 2025)	Decomposed binary checklist questions aggregated to ratings	Uses binary checklist questions, then reports aggregate score correlations.	BINARY The binary equivalence applies to raw yes/no checklist vectors, not aggregated scores.
5	MT-Bench / Chatbot Arena (Zheng et al., 2023)	Pairwise A/B/tie and single-answer ratings	Reports judge–human agreement, with and without ties.	PAIRWISE Relevant for tie handling; binary equivalence applies only after excluding or collapsing ties.
6	AlpacaEval / LC-AlpacaEval (Dubois et al., 2024)	Pairwise preference / win rate	Reports win-rate and length-controlled win-rate measures.	PAIRWISE No redundant correlation coefficients.
7	Auto-J (Li et al., 2023)	Pairwise A/B/tie plus rating and critique	Pairwise preference can include ties.	PAIRWISE Tie handling is explicit, but not framed as selective prediction or CA.
8	PandaLM (Wang et al., 2023b)	Pairwise preference judging	Reports F_1-style agreement relative to stronger judges or human preferences.	PAIRWISE No redundant Pearson/Spearman/Kendall coefficients found.
9	JudgeLM (Zhu et al., 2023)	Pairwise and classification-like judge settings	Reports agreement/accuracy-style measures.	PAIRWISE No redundant correlation coefficients found.
10	LLMBar (Zeng et al., 2023)	Pairwise instruction-following contrast	Evaluator chooses the instruction-following output from a pair.	PAIRWISE Motivates scale-aware metrics.
11	Prometheus (Kim et al., 2023)	Rubric-based direct assessment and pairwise evaluation	Pearson on graded rubric scores; accuracy on pairwise benchmarks.	GRADED Mixed setting; direct assessment and binary/pairwise validation are distinct.
12	Prometheus 2 (Kim et al., 2024)	Direct assessment plus pairwise ranking	Correlations for direct-assessment scores; agreement/accuracy for pairwise settings.	GRADED Boundary case: multiple correlations are mostly on graded scores.
13	FLASK (Ye et al., 2024)	Fine-grained 1–5 skill scores	Reports Pearson/Spearman/Kendall correlations.	GRADED Multi-correlation example, but not binary.
14	LLM-Rubric (Hashemi et al., 2024)	Rubric-question response distributions and 1–4 satisfaction prediction	Uses rubric dimensions and predicts overall human satisfaction.	GRADED Not a clean binary verdict setting in the reported overall evaluation.
15	ARES (Saad-Falcon et al., 2023)	RAG judge classifiers for context relevance, faithfulness, and relevance	Uses classifier-style evaluation and PPI confidence machinery.	BINARY Confusion-matrix and precision/recall reporting matter.
16	RAGAS (Es et al., 2023)	RAG faithfulness/relevance metrics	Reports agreement/validation-style quantities for RAG dimensions.	BINARY "Insufficient information" appears in prompts, but not as a reported CA estimand.
17	PoLL (Verga et al., 2024)	Panel/jury of LLM evaluators	Evaluates panels across multiple judge settings and datasets.	ENSEMBLE Multi-judge motivation; does not show panels as a production default.
18	ChatEval (Chan et al., 2023)	Multi-agent referee team	Multi-agent discussion followed by scoring/judgment.	ENSEMBLE Scale depends on task.
19	JUDGE-BENCH (Bavaresco et al., 2025)	Multiple datasets with categorical and graded annotations	Cohen's κ for categorical datasets, Spearman for graded datasets, Krippendorff's α for human agreement.	MULTI-CLASS Positive example of scale-aware metric selection.
20	FineSurE (Song et al., 2024)	Fine-grained summarization dimensions	Multi-dimensional summarization evaluation.	GRADED More relevant to scale/criteria framing than to binary redundancy.
21	TIGERScore (Jiang et al., 2023)	Trained explainable text-generation metric	Correlation-style meta-evaluation with human ratings.	GRADED Not binary.
22	InstructScore (Xu et al., 2023)	Score plus diagnostic feedback	Correlates generated scores with human ratings.	GRADED Not binary.
23	InstructEval (Chia et al., 2023)	Holistic evaluation of instruction-tuned LLMs	Includes Likert-style writing evaluation and task metrics.	GRADED Mostly not a binary judge-verdict setting.
24	JuStRank (Gera et al., 2024)	Pairwise/system-ranking judgments	Studies LLM judges as system rankers.	PAIRWISE Relevant especially after aggregation; no redundant correlation coefficients found.

Table A1: Audit of 24 recent LLM-as-judge and LLM-evaluator papers, coded by judged output scale, reported metrics, and relevance to the binary equivalence developed here. The accent tag in the last column marks each system's primary relevance category (BINARY, PAIRWISE, GRADED, MULTI-CLASS, or ENSEMBLE).

at $(0.4, 0.1, 0.2, 0.3)$ ($\pi = 0.5$, $\hat{\pi} = 0.6$, $\kappa = 0.400$, $\phi = 1/\sqrt{6} \approx 0.408$), the delta method gives $\mathrm{Var}(\hat{\kappa}) = 504/(625N) \approx 0.806/N$ and $\mathrm{Var}(\hat{\phi}) = 13/(16N) = 0.813/N$. The variance gap is second-order in marginal mismatch, mirroring the pointwise gap.