Correlation Is Not Agreement: Rethinking LLM Judge Validation

Anonymous ACL submission

Abstract

Validating LLM judges via correlation with human scores is insufficient for deployment decisions. Correlation measures ranking consistency; deployment requires *agreement*—whether a judge can substitute for a human rater at the segment level. We formalize this distinction through *Alpha Parity*, a three-stage framework combining scale-adaptive Krippendorff's α, non-inferiority testing against a task-adaptive margin, and a discriminative validity check against a random baseline. We assemble the largest per-annotator NLG evaluation corpus we are aware of—3.06M judgments across four MT annotation protocols (DA, DA+SQM, ESA, MQM) drawn from WMT 2022–2024 academic campaigns, Google MQM 2020–2024, and ToShip 2018–21, augmented by SummEval and QAGS—and find that 91% of 96 measurement configurations fall below accepted reliability thresholds. We apply Alpha Parity to eight frontier LLM judges from three model families and 188 WMT metric submissions. No single correlation threshold reproduces agreement-based verdicts across domains: a threshold that correctly separates judges on factual consistency (QAGS) produces 41 unsafe deployments on MT (WMT MQM); raising it eliminates those errors but rejects all 14 passing QAGS judges. Alpha Parity replaces threshold selection with a statistically grounded, per-configuration verdict that requires no externally imposed cutoff.

1 Introduction

Large language models (LLMs) are increasingly used as automated evaluators—LLM-as-a-Judge—across natural language generation (NLG) tasks, from machine translation (MT) quality estimation (Kocmi and Federmann, 2023) to summarization assessment (Liu et al., 2023), dialogue evaluation (Zheng et al., 2023), and reinforcement learning from human feedback (RLHF) preference labeling (Ouyang et al., 2022). The appeal is clear: LLM judges are fast, cheap, and scale to any number of items. The question is whether they are *reliable*.

The standard validation methodology computes correlation—Pearson's r or Spearman's ρ—between LLM judge scores and human scores. If correlation is "high," the judge is declared suitable for deployment. This practice is widespread in MT metrics shared tasks (Freitag et al., 2023), LLM evaluation benchmarks (Zheng et al., 2023), and RLHF training pipelines. But what correlation threshold warrants deployment? No principled answer exists—correlation is used to *rank* metrics ("which metric correlates best?"), not to make pass/fail deployment decisions. The threshold, if applied at all, is chosen ad hoc.

Correlation is the wrong measure for this decision. Bland and Altman (1986) demonstrated this for clinical measurement comparison in 1986: two instruments can have perfect correlation while disagreeing systematically on every measurement. A judge that scores every translation 30 points too high has $r = 1.0$ with the gold standard but would make incorrect quality decisions on every segment. Correlation measures *ranking*; deployment requires *agreement*.

We present three contributions. First, a cross-domain agreement landscape built on a corpus of 3.06M per-annotator judgments across four MT protocols (DA, DA+SQM, ESA, MQM)[1] drawn from WMT 2022–2025 academic campaigns, Google MQM 2020–2024, and ToShip 2018–21 (Kocmi et al., 2021), together with the foundational human parity study (Hassan et al., 2018), SummEval (Fabbri et al., 2021), and QAGS (Wang et al., 2020)—96 configurations, 91% below the $\alpha \geq 0.667$ threshold recommended for content analysis (Krippendorff, 2004). Second,

[1] DA: Direct Assessment; DA+SQM: DA with Simplified Quality Metric marking; ESA: Error Span Annotation; MQM: Multidimensional Quality Metrics.

the Alpha Parity framework: a three-stage pipeline combining scale-adaptive α, non-inferiority testing against a task-adaptive margin δ calibrated from human leave-one-out variability, and a discriminative validity check against a random baseline, producing a structured pass/fail matrix rather than a single correlation number. Third, an empirical demonstration across eight frontier LLM judges and 188 WMT metric submissions, spanning three measurement scales. We find that correlation is a weak predictor of agreement: across 199 (judge $\times$ configuration) rows, the meta-correlation between Pearson r and α_{diff} is only $r_{\text{meta}} = 0.40$, and no single correlation threshold reproduces agreement-based verdicts across domains.

The framework is domain-agnostic and applies wherever humans produce evaluation labels and LLM judges are proposed as substitutes—including RLHF preference annotation, content safety classification, and code review. Complete source code and reproduction scripts will be released upon acceptance.

2 Background

2.1 Correlation vs. Agreement

Bland and Altman (1986) showed that Pearson's r between two measurement methods can be high even when they disagree systematically: correlation is invariant under linear transformations. Krippendorff's α (Krippendorff, 2004) is the chance-corrected agreement coefficient we adopt throughout; it accepts any number of raters, any measurement level (nominal, ordinal, interval, ratio), and missing data, and attains its maximum of 1 only when raters agree on absolute values. We follow Krippendorff's convention that $\alpha \geq 0.667$ supports tentative reliability, noting that almost no NLG evaluation configuration in our corpus meets it.

Proposition 1 (Correlation Invariance Under Systematic Bias)**.** *Let $h_1, \ldots, h_n$ be human scores and $j_i = a \cdot h_i + b + \epsilon_i$ be judge scores with $a > 0$, $b \neq 0$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then Pearson$(h, j) \to 1$ as $\sigma \to 0$, regardless of the values of a and b. In contrast, Krippendorff's α at the interval level attains its maximum at $(a, b, \sigma) = (1, 0, 0)$ and decreases strictly as $|a - 1|$ or $|b|$ increases. Proof in Appendix E.*

Correlation cannot detect offset bias, scale compression, or their combination—all preserve ranking while degrading absolute agreement—and provides no principled deployment threshold.

2.2 Non-Inferiority Testing

Let α_H denote human-only pool agreement, α_{H+J} agreement with a candidate judge included, and $\alpha_{\text{diff}} = \alpha_{H+J} - \alpha_H$. Krippendorff's α tells us *how much* agreement exists in a pool, but not whether adding a judge would *meaningfully degrade* it. A single cutoff "accept the judge if pool α stays above c" fails because the needed cutoff depends on task difficulty: a high-agreement configuration ($\alpha_H \approx 0.9$) tolerates substantial degradation; a low-agreement one ($\alpha_H \approx 0.4$) tolerates almost none.

Clinical trials use non-inferiority testing to determine whether a new treatment is "not meaningfully worse" than a standard (Piaggio et al., 2012). We adapt this construction: let $\text{CI}^{\gamma}_{\text{lower}}$ denote the γ-level lower bound of the bootstrap distribution of α_{diff} (procedure in Appendix F). The judge passes if

$$\text{CI}^{0.95}_{\text{lower}} > -\delta, \tag{1}$$

where δ is a pre-registered margin. We recommend $\delta = \max(2\sigma_{\text{LOO}}, 0.05)$, where σ_{LOO} is the leave-one-out standard deviation of α computed by removing each human rater in turn—making δ task-adaptive: strict for high-agreement tasks, lenient for noisy ones.

2.3 Related Work

Appendix I surveys the relevant literature: correlation critiques in MT evaluation (Kocmi et al., 2021, 2024b; Deutsch et al., 2023; Moghe et al., 2023; Perrella et al., 2024a; Proietti et al., 2025); concurrent judge-validation frameworks pairing correlation with other signals (Han et al., 2025; Song et al., 2026; Nasser et al., 2026; Choi et al., 2026); and clinical-measurement precedent for non-inferiority framing (Bland and Altman, 1986; Artstein and Poesio, 2008; Grolleau et al., 2025; Aali et al., 2025). Alpha Parity differs by combining scale-adaptive α, a task-adaptive δ derived from human leave-one-out variability, a discriminative random-baseline check, and cross-domain validation revealing construct-mismatch and quality-enhancer regimes.

3 The Alpha Parity Framework

Alpha Parity validates an LLM judge against a specific task configuration through a three-stage pipeline (Figure 1). The stages are strictly serial: each takes the output of the previous stage as input,

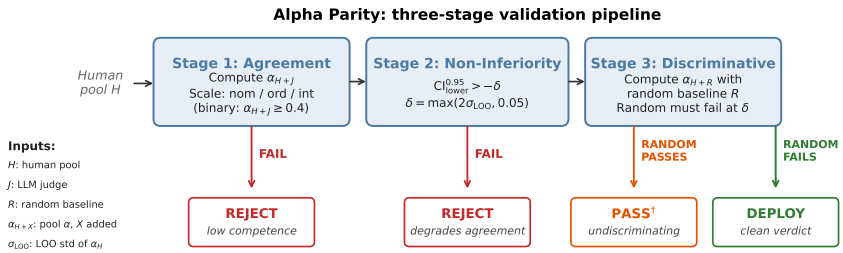

Figure 1: The Alpha Parity three-stage serial pipeline. Stage 1 measures agreement at the appropriate scale and, for binary tasks, enforces a competence gate ($\alpha_{H+J} \geq 0.4$). Stage 2 tests substitutability via bootstrap non-inferiority against the task-adaptive margin δ. Stage 3 re-runs Stage 2 on a distribution-matched random baseline: if random also passes, the verdict is flagged PASS†.

and failure at any stage terminates the evaluation with a FAIL verdict.

Stage 1: Agreement measurement. *Input*: judge predictions J and a human rater pool H. *Output*: the pool-agreement scalars α_H and α_{H+J}, at the measurement level appropriate to the task. Krippendorff's α uses a different distance function for nominal (binary), ordinal (Likert), interval (continuous DA or ESA), and ratio data; choosing the wrong level can change α by up to 0.3 on the same data (Krippendorff, 2004). For binary tasks only, Stage 1 additionally enforces a competence floor: $\alpha_{H+J} \geq 0.4$. This floor is well below Krippendorff's tentative-reliability threshold ($\alpha \geq 0.667$ (Krippendorff, 2004)) and is not itself a reliability claim—it is a sanity gate ruling out judges whose binary labels are grossly inconsistent with the pool, for which the downstream non-inferiority test would be uninformative. This gate exists because binary disagreement cannot be meaningfully averaged—a judge that flips 40% of labels has no use case, even if adding it barely perturbs pool α. For ordinal and interval tasks, Stage 1 computes α_H and α_{H+J} and passes them to Stage 2 without further checks.

Stage 2: Non-inferiority test. *Input*: α_H, α_{H+J}, and the bootstrap samples from Stage 1. *Output*: PASS or FAIL at pre-registered margin δ. Test $H_0\colon \alpha_{\text{diff}} < -\delta$ against $H_1\colon \alpha_{\text{diff}} \geq -\delta$, where $\alpha_{\text{diff}} = \alpha_{H+J} - \alpha_H$, using a bootstrap CI (10,000 resamples, block bootstrap at the document level). The task-adaptive margin $\delta = \max(2\sigma_{\text{LOO}}, 0.05)$ is calibrated from the leave-one-out standard deviation of the human pool. Worked example: on QAGS CNN/DM, $\alpha_H = 0.514$ and adding GPT-4o yields $\alpha_{H+J} = 0.569$, so $\alpha_{\text{diff}} = +0.055$; the bootstrap 95% lower CI on α_{diff} is -0.012, which exceeds $-\delta = -0.05$, so GPT-4o *passes*. For reporting, we also compute $\delta/2$ (stricter) and 2δ (more lenient) for sensitivity.

Stage 3: Discriminative validity check. *Input*: the Stage 2 test specification. *Output*: either confirms the Stage 2 verdict, or flags it as PASS† ("human baseline too noisy to discriminate"). Run the identical non-inferiority test on a distribution-matched random baseline R that has the same marginal score distribution as J but no signal; R's α with the human pool is not zero because z-normalization preserves each rater's marginal distribution, so baseline-rate agreement is structurally retained under random permutation (the "null of agreement explainable by score distribution alone"). If R also passes Stage 2, the human pool cannot distinguish real judges from noise at this margin, and any Stage 2 PASS at this configuration is annotated PASS†. Concretely, on SummEval relevance, expert $\sigma_{\text{LOO}} = 0.097$ produces $\delta = 0.19$; our random baseline passes with $\alpha_{\text{diff}} = -0.19$, so all eight real-judge passes are flagged. Stages 2 and 3 answer different statistical questions and must be reported separately, following clinical-trial practice; Appendix F gives the full procedure.

The framework produces a *pass/fail matrix*—one row per judge, one column per task configuration—which is the primary deliverable, not a single aggregated score.

4 Cross-Domain Agreement Landscape

The corpus comprises 3.06M MT judgments under four protocols (DA, DA+SQM, ESA, MQM), augmented by SummEval and QAGS. Together these yield 96 measurement configurations, 89 of which produce a valid Krippendorff's α at the cell level (Figure 2).

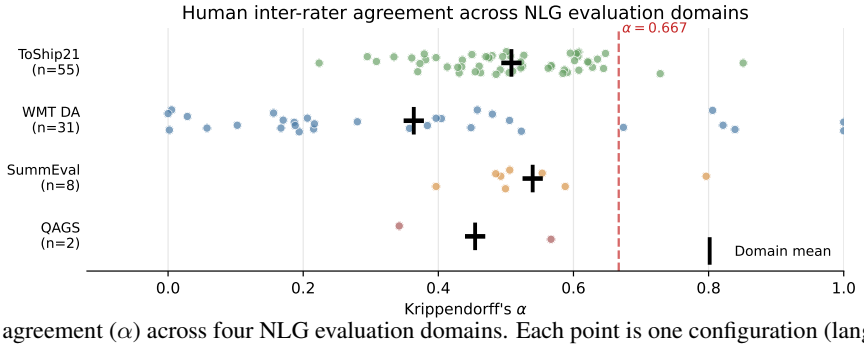

Figure 2: Human agreement (α) across four NLG evaluation domains. Each point is one configuration (language pair, quality dimension, or dataset); black markers are domain means. Only 9 of 96 configurations exceed the $\alpha = 0.667$ reliability threshold (Krippendorff, 2004).

Table 1: Per-protocol Krippendorff's α pooled across the MT IAA corpus. "Cells" counts (year $\times$ LP) cells contributing to each pool; α (L2) is computed on a 5K random sample of doubly-rated units (seed 42); α (L1) is the mean across per-cell estimates; "Median overlap" is the median fraction of units in a cell rated by ≥ 2 annotators. DA and DA+SQM are statistically indistinguishable; ESA's L2 reflects the high-overlap WMT24 cells (discussed below).

Protocol	Cells	α (L2)	α (L1)	Median overlap
DA	56	0.418	0.46	0.000
DA+SQM	21	0.424	0.47	0.373
ESA	9	0.826	0.37	0.001
MQM	3	0.523	0.37	0.000

4.1 Machine Translation

The MT corpus combines public per-annotator releases: WMT22–24 academic campaigns under DA+SQM and ESA (Kocmi et al., 2024c,d), Google MQM 2020–2024 (Freitag et al., 2021), ToShip 2018–21 (Kocmi et al., 2021), and WMT-DA legacy items where per-annotator data is publicly accessible. Coverage concentrates in 2020 and 2022–2024; pre-2022 academic per-annotator DA is not currently consolidated in public mirrors (see §7).

Per-protocol agreement. Table 1 reports α pooled across all cells of each protocol, computed on a fixed 5K random sample of doubly-rated units per pool (necessary because the `simpledorff` implementation scales as $O(n \cdot m^2)$ and the full DA pool exceeds 1.85M units). Three findings stand out. First, **DA and DA+SQM are statistically indistinguishable** at the protocol level (0.418 vs. 0.424); the L1 cell-mean numbers (0.46 vs. 0.47, $n = 56$ and 21 cells) confirm this is not an artefact of the L2 sample. Despite different annotation interfaces (single 0–100 slider vs. a two-pass scoring with quality marking), per-protocol agreement is equivalent. Second, **ESA reaches** $\alpha = 0.826$—substantially higher than any scoring protocol—when designed for multi-annotator overlap. Third, **MQM sits between DA and ESA at** $\alpha = 0.523$: the error-typology constraint structures annotator behaviour more than free scoring but less than full span annotation.

Overlap structure as a hidden axis. Per-protocol α is computed only over the doubly-rated subset of each cell. Median per-cell overlap varies dramatically across protocols (DA 0.000, DA+SQM 0.373, ESA 0.001, MQM 0.000): most cells are by design single-annotator-per-segment, and α is estimable only on the overlapping subset. WMT24 ESA is the sharpest case: four cells (en-zh, en-ja, en-cs, en-hi) with 42–90% overlap give $\alpha \in [0.70, 0.85]$; five cells (cs-uk, en-ru, en-is, en-uk, en-es) with $\leq 0.1\%$ overlap give $\alpha \in [0.0005, 0.082]$, with bootstrap CIs cleanly separating the two regimes at the 95% level. Same protocol, same year, radically different agreement—a campaign-design property, not a protocol-intrinsic one. §3 uses this stratification as a stress test for uniform-threshold validation.

Statistical headroom at industrial scale. ToShip cells with $\geq$ 15K judgments and $\geq$ 100 annotators routinely produce bootstrap CIs of width 0.025–0.05 (e.g., `2020 en-hu DA`: $\alpha = 0.563$, CI $= [0.551, 0.576]$, $n = 17{,}241$). At these volumes the $\delta = 0.05$ margin is comfortably resolvable. 12 of 28 completed DA cells fall in $\alpha \in [0.40, 0.55]$ with overlapping CIs, supporting an "$\alpha \approx 0.45$–0.50 floor for industrial DA at scale" summary; counter-examples exist at the same scale (`2022 fr-en DA`, $\alpha = 0.27$ at 67K units; Appendix B.4), anchoring the need for per-cell rather than global calibration.

4.2 Summarization and Factual QA

SummEval (Fabbri et al., 2021) expert ratings reach $\alpha = 0.90$ on consistency—the only measurement in any domain clearly exceeding 0.667; expert coherence is 0.73, fluency and relevance lower (0.56, 0.45). Crowd SummEval clusters at $\alpha \approx 0.48$–0.51 regardless of dimension. QAGS (Wang et al., 2020) gives $\alpha = 0.51$ on CNN/DM (162 crowd workers, 714 sentences) and 0.34 on XSum (84 workers, 239 sentences).

4.3 Cross-Domain Summary

Across all 96 configurations and 89 valid cells, 87 fall below $\alpha = 0.667$ (91%; median $\alpha \approx 0.49$; distribution in Appendix A). The headline pattern is that overlap structure—a property of campaign instrumentation, not of the protocol or task—can dominate raw α comparisons. Any framework that validates judges against human labels must therefore (a) report agreement conditional on the same overlap stratum that produces the human reference, and (b) compare judges to a margin calibrated from local human variability rather than a fixed cutoff.

5 Demonstrations

We demonstrate Alpha Parity on 188 WMT metric submissions (§5.1.1), eight LLM judges on QAGS binary factual consistency (§5.1.2), SummEval ordinal quality (§5.1.3), and WMT25 ESA continuous translation quality (§5.1.5); adversarial bias-injection results are in Appendix H, and Appendix C retrospectively applies Alpha Parity to the human parity claim of Hassan et al. (2018).

5.1 LLM Judge Validation

The WMT metrics analysis (§5.1.1) establishes the framework on segment-level MT using 188 existing metric submissions; we then test whether it generalizes across measurement scales by evaluating LLM judges on binary factual consistency (§5.1.2, QAGS), ordinal summary quality (§5.1.3, SummEval), and returning to continuous MT under the ESA protocol (§5.1.5, WMT25).

5.1.1 WMT Metrics as Pseudo-Judges

As a first application of Alpha Parity, we treated every metric submitted to the WMT22 and WMT23 Metrics shared tasks as a pseudo-judge—extending Kocmi et al. (Kocmi et al., 2021) from system-level pairwise accuracy to segment-level agreement. We computed α (with the metric added to the human MQM rater pool) and Pearson correlation against human mean scores, producing 188 metric×language pair combinations at zero cost. Of 14 testset×LP combinations in `mt-metrics-eval`, only three have sufficient rater overlap for segment-level α: WMT23 en-de ($\alpha_H = 0.644$), WMT23 zh-en (0.577), WMT22 en-de (0.466). WMT24 data had zero items with two or more raters—itself a finding: *the majority of WMT evaluation campaigns lack the annotation design needed for agreement-based validation.*

Results. Only 11 of 188 combinations (5.9%) pass at $\delta = 0.05$. All passes occur on WMT23 zh-en (10) or WMT22 en-de (1); no metric passes on WMT23 en-de, where human agreement is highest. MetricX-23 (Freitag et al., 2023), and its successor MetricX-24 (Juraska et al., 2024), are the only families that *increase* pool α (best variant $\alpha_{\text{diff}} = +0.027$ on zh-en). String-matching metrics destroy agreement: BLEU reduces α by 0.76, chrF by 0.89 on WMT23 en-de. GEMBA-MQM (Kocmi and Federmann, 2023b) achieves the highest correlation among LLM submissions on en-de ($r = 0.50$) but fails α on both LPs ($\alpha_{\text{diff}} = -0.087$ on en-de, -0.151 on zh-en)—state-of-the-art LLM judges do not yet match expert MQM segment agreement.

Bidirectional divergence. Correlation and agreement diverge in two directions: *UNSAFE* (correlation suggests deployment, α rejects) and *MISSED* (correlation suggests rejection, α passes). Sweeping all thresholds (Figure 3): at $r > 0.3$, correlation would deploy 112 metrics that degrade human agreement; at $r > 0.70$ and above, UNSAFE vanishes but all 11 α-passing metrics are rejected. No correlation threshold eliminates both error types simultaneously. Across all measurements (188 WMT + 16 QAGS), meta-correlation between Pearson r and α_{diff} is only $r_{\text{meta}} = 0.54$[2]—correlation explains less than 30% of the variance in agreement impact (scatter in Appendix A.1).

Recommendation. MetricX-23 and MetricX-24 are the only learned metric families that increase pool α on MQM-annotated data (best $\alpha_{\text{diff}} = +0.027$); string-matching metrics (BLEU, chrF) destroy agreement and should not be used as proxies

[2]Unweighted Pearson over the 204 (metric/judge × configuration) rows. Item-count-weighted variants on the same pool give $r_{\text{meta}} \in [0.47, 0.52]$ (linear n, $\sqrt{n}$, $\log_{10} n$); WMT cells contribute 99.7% of mass under linear weighting, so the unweighted estimate is the more pool-comparable one.

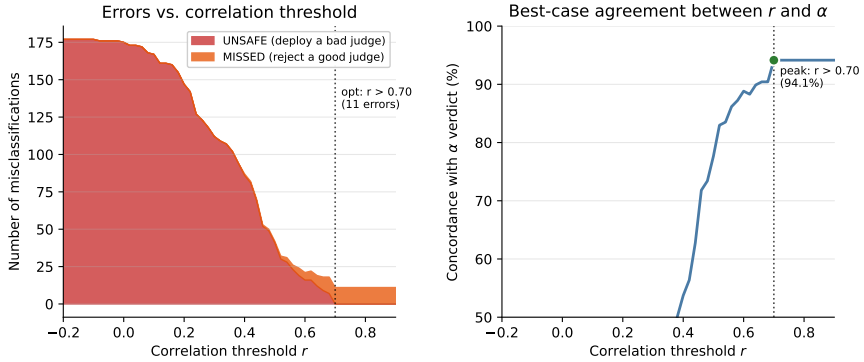

Figure 3: Errors at each correlation threshold across the 188 WMT metric submissions, treating α non-inferiority as ground truth. Left: at $r > 0.3$, 112 UNSAFE cases arise; the optimal threshold ($r > 0.70$) eliminates UNSAFE but produces 11 MISSED. Right: peak concordance is 94.1%.

Table 2: QAGS judge results. $\alpha_H = 0.514$ (CNN/DM) and 0.342 (XSum). Seven of eight LLM judges improve human α when added to the pool (positive α_{diff}) and pass the non-inferiority test, yet all have $r < 0.8$ and would be rejected by correlation-based validation. Gemini Pro is the sole failure.

Subset	Judge	r	α_{diff}	Verdict
CNN/DM	GPT-4o	0.760	**+0.055**	PASS
	Claude Haiku	0.743	**+0.049**	PASS
	Claude Sonnet	0.741	**+0.046**	PASS
	GPT-4o-mini	0.722	**+0.040**	PASS
	o3-mini	0.709	**+0.030**	PASS
	Claude Opus	0.697	**+0.028**	PASS
	Gemini Flash	0.705	**+0.025**	PASS
	Gemini Pro	0.441	−0.131	FAIL
	Random	−0.018	−0.275	FAIL
XSum	Claude Opus	0.661	**+0.076**	PASS
	o3-mini	0.661	**+0.077**	PASS
	Claude Haiku	0.648	**+0.071**	PASS
	Claude Sonnet	0.648	**+0.071**	PASS
	GPT-4o-mini	0.638	**+0.068**	PASS
	GPT-4o	0.631	**+0.051**	PASS
	Gemini Flash	0.561	**+0.030**	PASS
	Gemini Pro	0.218	−0.166	FAIL
	Random	−0.083	−0.209	FAIL

for segment-level quality.

5.1.2 QAGS: Factual Consistency

Having established the framework on segment-level MT metrics (§5.1.1, interval scale), we now test whether it generalizes to binary labels. QAGS offers two properties MT annotation lacks—expert-adjudicated ground-truth labels and large crowd pools (162 on CNN/DM, 84 on XSum)—letting us evaluate judges against a baseline where both signal and noise are known. Eight LLM judges from three families rated 953 summary sentences (CNN/DM: 714; XSum: 239) with a binary prompt ("Is this sentence factually consistent? YES/NO") (Wang et al., 2020).

Table 2 reveals 14 MISSED divergence cases: every passing LLM judge *increases* pool α when added, yet has $r < 0.8$—correlation-based validation would reject all 14. Rankings shuffle between subsets (GPT-4o 1st on CNN/DM, 6th on XSum; Opus the reverse), so model scale provides no reliable advantage on binary tasks; o3-mini helps more on the harder XSum subset. Judges help more where humans disagree more: on XSum ($\alpha_H = 0.342$) improvements are $+0.030$ to $+0.077$, while on CNN/DM ($\alpha_H = 0.514$) they are $+0.025$ to $+0.055$. A consistent LLM judge stabilizes a noisy human pool—exactly what agreement-based validation detects and correlation cannot.

Recommendation. GPT-4o is the strongest single choice on CNN/DM ($\alpha_{\text{diff}} = +0.055$); Opus and o3-mini tie on XSum ($\alpha_{\text{diff}} = +0.076/+0.077$). Seven of eight LLM judges pass; only Gemini Pro fails both subsets.

5.1.3 SummEval: Ordinal Scale Validation

We extended the framework to SummEval (Fabbri et al., 2021), where judges rate summaries on four dimensions (coherence, consistency, fluency, relevance) using a 1–5 ordinal scale. **Coherence**: seven of eight pass; Opus ($r = 0.60$, $\alpha_{\text{diff}} = -0.029$) and o3-mini ($r = 0.54$, -0.047) degrade least; Flash ($r = 0.35$, -0.147) passes despite the lowest correlation. **Consistency**: every judge fails, including Opus ($r = 0.73$, $\alpha_{\text{diff}} = -0.138$). Notably, Opus's consistency α_{diff} is comparable in magnitude to Flash's coherence (-0.147), yet the verdicts are opposite because the adaptive margin differs ($\delta = 0.05$ vs. 0.21)—a fixed threshold cannot distinguish these cases. **Relevance**: all eight real judges pass at $\delta = 0.19$. Our random baseline also passes ($\alpha_{\text{diff}} = -0.19$), indicating the human pool is too noisy ($\sigma_{\text{LOO}} = 0.097$) to discriminate;

all eight passes are flagged † per Stage 3. **Fluency**: no judge achieves acceptable agreement—likely poorly defined for LLM evaluation or requiring dimension-specific prompting.

Recommendation. Claude Opus is the strongest judge for coherence ($\alpha_{\text{diff}} = -0.029$, within margin); no judge achieves non-inferiority on consistency or fluency and these dimensions still require human raters; relevance passes carry Stage 3 PASS† flags because the expert pool is too noisy to discriminate real judges from the random baseline.

5.1.4 MQM: Construct Mismatch

On MQM-annotated data, all eight real judges fail catastrophically ($\alpha_{\text{diff}} \approx -0.82$ to -0.89): DA-style holistic scores and MQM error counts measure different properties of translation quality. The next subsection (WMT25 ESA) confirms the interpretation—the same judges pass on ESA-annotated data for overlapping language families.

5.1.5 ESA: Continuous Translation Quality

WMT25's ESA protocol uses continuous quality scores, sidestepping the MQM construct mismatch of §5.1.4. We evaluated 12 language pairs spanning high-, mid-, and low-resource settings.

43 passes across 13 language pairs. Claude Opus passes on 10 LPs (cs-de, en-cs, en-et, en-is, en-it, en-ja, en-mas, en-ru, en-sr, en-zh), o3-mini on 8 (cs-de, cs-uk, en-bho, en-mas, en-ru, en-sr, en-uk, en-zh), Gemini Flash on 7 diverse LPs (including Bhojpuri and Maasai), Claude Sonnet on 6, Claude Haiku, Gemini Pro, GPT-4o, and GPT-4o-mini each on 3 low-resource LPs (en-mas, en-ru, en-sr) where human baselines admit them ($\delta = 0.55$ on en-mas). Only en-ar admits no real-judge pass. Several judges *improve* human agreement: Flash on en-mas ($\alpha_{\text{diff}} = +0.25$, the largest gain in our data), Opus and GPT-4o on en-mas ($+0.22$), Pro on en-mas ($+0.21$), Opus on cs-de ($+0.09$). The "quality enhancer" pattern from QAGS extends to MT, with the largest gains on low-resource pairs where human annotation is scarcest. Per-LP routing requires per-task validation; no single correlation threshold captures this specialization.

5.1.6 Failure-Mode Taxonomy

Three failure modes emerge across the four configurations: *low competence* (Gemini Pro on QAGS, $r = 0.44$, $\alpha_{\text{diff}} = -0.131$), *construct mismatch* (all judges on WMT23 MQM, $\alpha_{\text{diff}} \approx -0.8$; holistic scores vs. discrete error counts), and *latent scale mismatch* (Opus on SummEval consistency, $r = 0.73$, $\alpha_{\text{diff}} = -0.138$; ranks items similarly but disagrees on absolute quality). Only the last is invisible to correlation-based validation.

5.1.7 Task-Adaptive Margins in Action

The margin $\delta = \max(2\sigma_{\text{LOO}}, 0.05)$ ranges from the 0.05 floor (QAGS CNN/DM, SummEval consistency, WMT23 en-de MQM, WMT25 cs-de) to $\delta = 0.548$ on WMT25 en-mas where only four raters cover a very low-resource pair. These values explain why Opus's $\alpha_{\text{diff}} = -0.138$ on SummEval consistency (FAIL, $\delta = 0.05$) carries a comparable magnitude to Flash's -0.147 on coherence (PASS, $\delta = 0.21$): the adaptive margin differs. The full δ-by-configuration table is in Appendix A.2.

5.1.8 No Universal Threshold

A threshold analysis across the 199 judge×configuration rows (Appendix A.4) yields per-domain optima spanning $[0.35, 0.50]$; no single global threshold simultaneously achieves all four. Meta-correlation is $r_{\text{meta}} = 0.40$—correlation explains only 16% of variance in agreement impact. Alpha Parity eliminates threshold selection: each judge×task receives a statistically grounded verdict.

Practitioner recommendations. Per-configuration recommendations are given in §5.1.1–§5.1.5. The broader recommendation is not a specific judge but the per-configuration verdict: deploy where the judge passes, retain humans where it fails, and refresh verdicts when the judge or annotation protocol changes. A four-injection adversarial analysis on GPT-4o confirms that Bias A (scale compression) flips the Alpha Parity verdict from PASS to FAIL while leaving Pearson r provably invariant (Appendix H).

6 Implications

The landscape reveals human agreement on segment-level quality is moderate at best: across 89 valid cells spanning four MT protocols, summarization, and factual QA, only 9 exceed Krippendorff's $\alpha = 0.667$ threshold, and the DA pool averages $\alpha \approx 0.42$ even at industrial scale ($\geq$ 15K judgments, $\geq$ 100 annotators per cell). Researchers should report the human baseline alongside judge performance; the meta-correlation $r_{\text{meta}} = 0.40$

between r and α_{diff} (Figure 4) shows correlation-based rankings do not reliably predict which metrics maintain human agreement. The pass/fail matrix supports targeted automation and per-task routing: deploy where the judge passes, retain humans where it fails. Validation is also not permanent—when annotation quality improves (WMT23 mean $\alpha = 0.42 \rightarrow$ WMT24 ESA α up to 0.84), a previously passing judge may fail re-validation because the human ceiling has risen. Alpha Parity adapts automatically (the non-inferiority test compares against the *current* human baseline); correlation-based validation does not.

Quality enhancers vs. tolerable substitutes. A striking pattern emerges: on QAGS binary tasks with large crowd-worker pools (162 raters), all passing judges have *positive* α_{diff}—they improve human agreement when added to the pool. On SummEval ordinal tasks with small expert pools (3 raters), most passing judges have *negative* α_{diff} but within the margin δ, as non-inferiority testing explicitly permits. LLM judges are not merely cheap substitutes—in noisy annotation settings they are *quality enhancers* that increase the reliability of the resulting dataset. Practitioners should consult the sensitivity analysis ($\delta/2$, δ, 2δ); Appendix D discusses implications for RLHF reward modeling, safety evaluation, and auditability requirements.

7 Limitations

Shared annotator bias. If all raters share a systematic blind spot, an LLM judge replicating this bias will pass. Like Bland and Altman's method comparison, Alpha Parity measures whether a judge can substitute for a human rater, not whether the human raters are correct.

Sparse annotation overlap. With fewer than three raters per item, α estimates have high variance; we require bootstrap CIs and recommend ≥ 5 raters per item. The Stage 3 random-baseline check flags this regime: when random passes (e.g., SummEval relevance, $\alpha_{\text{diff}} = -0.19$ at $\delta = 0.19$), the verdict means "the human baseline is too noisy to discriminate" rather than "the judge is deployable"—practitioners should supplement with an absolute quality floor in such settings.

Skewed quality distributions. When the vast majority of items are high-quality, α may be inflated by agreement on the majority class. We recommend supplementary analysis on the error subset.

Snapshot verdicts. A PASS verdict reflects the judge's behavior at evaluation time. LLM API providers update models silently; future work should extend Alpha Parity to continuous monitoring (e.g., periodic re-validation against a refreshing canary set) as data and model distributions shift.

Proof scope vs. framework scope. Proposition 1 uses the two-rater closed form of Krippendorff's α, the concordance correlation coefficient (Lin, 1989), for analytic tractability. The framework operates on pools of arbitrary size (162 raters on QAGS CNN/DM, 3 on SummEval, 4–12 on WMT25); the proof's core mechanism—α responds to absolute distances while r is affine-invariant—generalizes to multi-rater α via the underlying pairwise-distance formulation.

Public per-annotator coverage. Per-annotator data for WMT 2017–2021 academic DA is not currently consolidated in public mirrors, so the DA pool is dominated (76% of rows) by ToShip 2018–21 commercial campaigns. ToShip text is additionally hashed in public releases, so item-level diagnostics require unhashed access.

No YouTube-derived speech subsets. Due to access restrictions imposed by the YouTube platform, YouTube-sourced content was not included in the analysis and represents a limitation of the current work.

8 Conclusion

Correlation is not agreement. Validating LLM judges via correlation conflates ranking consistency with measurement agreement, masking systematic biases that agreement-based analysis detects. Alpha Parity provides the missing framework: per-judge, per-task validation with a statistical test, a declared margin, and an auditable result. Our cross-domain landscape shows 91% of 96 configurations fall below accepted reliability thresholds; across eight frontier LLM judges and 188 WMT metric submissions, correlation is a weak predictor of agreement ($r_{\text{meta}} = 0.40$) and no single threshold reproduces agreement-based verdicts. We will release the Alpha Parity toolkit as open-source upon acceptance.

Ethical Considerations

This work is a measurement methodology consuming existing human-evaluation data; we collect no new human-subject data, and all corpora used (WMT 2022–2024 academic, Google MQM

2020–2024, ToShip 2018–21, SummEval, QAGS, Hassan et al.) are public releases with rater identities anonymized at source. The framework's main societal effect is to raise the evidentiary bar for deploying LLM judges in roles previously held by human annotators, which we argue is protective: it surfaces failure modes (scale compression, construct mismatch, latent bias) that correlation-based validation misses. One dual-use risk worth naming: an adversary who knew the margin δ could fine-tune a judge to sit just inside it; the Stage 3 random-baseline check partially mitigates this but does not eliminate the residual incentive.

References

Bland, J. M. and Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986.

Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 2nd edition, 2004.

Kocmi, T. and Federmann, C. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of EAMT*, 2023.

Kocmi, T. and Federmann, C. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of WMT*, pages 768–775, 2023.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of NeurIPS*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*, 2022.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. Results of WMT23 metrics shared task. In *Proceedings of WMT*, 2023.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*, 2018.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:568–600, 2021.

Wang, A., Cho, K., and Lewis, M. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of ACL*, pages 5008–5020, 2020.

Piaggio, G., Elbourne, D. R., Pocock, S. J., Evans, S. J., and Altman, D. G. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*, 308(24):2594–2604, 2012.

Juraska, J., Deutsch, D., Finkelstein, M., and Freitag, M. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of WMT*, 2024.

Lavie, A., Hanneman, G., Agrawal, S., et al. Findings of the WMT25 shared task on automated translation evaluation systems. In *Proceedings of WMT*, pages 436–483, 2025.

Rei, R., Guerreiro, N. M., Pombal, J., Alves, J., et al. Tower+: Bridging generality and translation specialization in multilingual LLMs. *arXiv preprint arXiv:2506.17080*, 2025.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., and Menezes, A. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of WMT*, pages 478–494, 2021.

Kocmi, T., Zouhar, V., Federmann, C., and Post, M. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of ACL*, pages 1999–2014, 2024.

Läubli, S., Sennrich, R., and Volk, M. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of EMNLP*, pages 4791–4796, 2018.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Federmann, C., et al. Findings of the WMT24 general machine translation shared task. In *Proceedings of WMT*, pages 1–46, 2024.

Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., et al. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of WMT*, pages 1440–1453, 2024.

Bland, J. M. and Altman, D. G. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, 1999.

Krippendorff, K. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*, 2011.

Artstein, R. and Poesio, M. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Wellek, S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, 2nd edition, 2010.

Hayes, A. F. and Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021.

Freitag, M., Rei, R., Mathur, N., Lo, C., et al. Results of WMT22 metrics shared task: Stop using BLEU—neural metrics are better and more robust. In *Proceedings of WMT*, 2022.

Kim, S., Shin, J., Cho, Y., et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *Proceedings of ICLR*, 2024.

Gu, J., Ye, X., Gao, Z., et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Wang, P., Li, L., Chen, L., et al. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Panickssery, A., Bowman, S. R., and Feng, S. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Stureborg, R., Alikaniotis, D., and Suhara, Y. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*, pages 2685–2702, 2020.

Han, S., Titericz Junior, G., Balough, T., and Zhou, W. Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement. *arXiv preprint arXiv:2510.09738*, 2025.

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

Nasser, W., et al. Evaluative fingerprints: Stable and systematic differences in LLM evaluator behavior. *arXiv preprint arXiv:2601.05114*, 2026.

Choi, J., Park, S., Cho, C., Park, H., and Kim, B. Diagnosing the reliability of LLM-as-a-judge via item response theory. *arXiv preprint arXiv:2602.00521*, 2026.

Leiter, C. and Eger, S. Neither valid nor reliable? Investigating the use of LLMs as judges. *arXiv preprint arXiv:2508.18076*, 2025.

Schaeffer, R., et al. Grading scale impact on LLM-as-a-judge: Human-LLM alignment is highest on 0–5 grading scale. *arXiv preprint arXiv:2601.03444*, 2026.

Thakur, A. S., et al. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. *arXiv preprint arXiv:2510.27106*, 2025.

Rosso, G. A note on the correlation coefficient and Bland–Altman analysis in medical research. *Radiology*, 2015.

Edwards, D., et al. Pearson's correlation coefficient: An inadequate measure of agreement in sonography. *Journal of Diagnostic Medical Sonography*, 2021.

Landesberg, E. When LLM judge scores look good but best-of-n decisions fail. *arXiv preprint arXiv:2603.12520*, 2026.

Shen, M., Balashankar, A., Fisch, A., Madras, D., and Rodrigues, M. Robust LLM performance certification via constrained maximum likelihood estimation. *arXiv preprint arXiv:2604.03257*, 2026.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Deutsch, D., Foster, G., and Freitag, M. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of EMNLP 2023*, pages 12914–12929, 2023.

Grolleau, F., Alsentzer, E., Keyes, T., Chung, P., Swaminathan, A., Aali, A., Hom, J., Huynh, T., Lew, T., Liang, A. S., et al. MedFactEval and MedAgentBrief: A framework and workflow for generating and evaluating factual clinical summaries. *arXiv preprint arXiv:2509.05878*, 2025.

Moghe, N., Fazla, A., Amrhein, C., Kocmi, T., Steedman, M., Birch, A., Sennrich, R., and Guillou, L. Machine translation meta evaluation through translation accuracy challenge sets (ACES). *arXiv preprint arXiv:2401.16313*, 2024.

Kovacs, G., Deutsch, D., and Freitag, M. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, 2024.

Zouhar, V. et al. Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models. *arXiv preprint arXiv:2505.01761*, 2025.

Cui, H. LLMs are not scorers: Rethinking MT evaluation with generation-based methods. *arXiv preprint arXiv:2505.16129*, 2025.

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F. T., Neubig, G., Garg, A., Clark, J. H., Freitag, M., and Firat, O. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, 2023.

Tan, S. and Monz, C. ReMedy: Learning machine translation evaluation from human preferences with reward modeling. *arXiv preprint arXiv:2504.13630*, 2025.

Finkelstein, M. et al. From jack of all trades to master of one: Specializing LLM-based autoraters to a test set. *arXiv preprint arXiv:2411.15387*, 2024.

Sandan, I. B., Dinh, T. A., and Niehues, J. Knockout LLM assessment: Using large language models for evaluations through iterative pairwise comparisons. In *Proceedings of GEM @ ACL 2025*, 2025.

Wang, S., Arnoult, S., and Kamran, A. Large language models as annotators for machine translation quality estimation. *arXiv preprint arXiv:2603.10775*, 2026.

Eigler, L., Libovický, J., and Hurych, D. LLM as a meta-judge: Synthetic data for NLP evaluation metric validation. *arXiv preprint arXiv:2603.09403*, 2026.

Perrella, S., Proietti, L., Scirè, A., Barba, E., and Navigli, R. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, 2024.

Perrella, S., Proietti, L., Huguet Cabot, P.-L., Barba, E., and Navigli, R. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, 2024.

DiIanni, C. and Deutsch, D. Don't sweat the small stuff: Segment-level meta-evaluation based on pairwise difference correlation. *arXiv preprint arXiv:2509.25546*, 2025.

Yamagishi, Y., Takamatsu, A., Hamaguchi, Y., Kikuchi, T., Hanaoka, S., Yoshikawa, T., and Abe, O. Blinded radiologist and LLM-based evaluation of LLM-generated Japanese translations of chest CT reports: Comparative study. *arXiv preprint arXiv:2604.02207*, 2026.

Flamich, G., Vilar, D., Peter, J.-T., and Freitag, M. You cannot feed two birds with one score: the accuracy-naturalness tradeoff in translation. *arXiv preprint arXiv:2503.24013*, 2025.

Zhang, R., Zhao, W., and Eger, S. How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, 2025.

Lin, L. I.-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989.

Proietti, L., Perrella, S., and Navigli, R. Has machine translation evaluation achieved human parity? The human reference and the limits of progress. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–813, 2025.

Son, G., Yoon, D., Suk, J., Aula-Blasco, J., Aslan, M., Kim, V. T., Islam, S. B., Prats-Cristà, J., Tormo-Bañuelos, L., and Kim, S. MM-Eval: A multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models. *arXiv preprint arXiv:2410.17578*, 2024.

Zhan, R., Huang, Z., Yang, X., Chao, L. S., Yang, M., and Wong, D. F. Are large reasoning models good translation evaluators? Analysis and performance boost. *arXiv preprint arXiv:2510.20780*, 2025.

Moghe, N., Sherborne, T., Steedman, M., and Birch, A. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, 2023.

Deguchi, H., Chousa, K., and Sakai, Y. Hacking neural evaluation metrics with single hub text. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2026.

Aali, A., Bikia, V., Varma, M., Chiou, N., Ostmeier, S., Singhvi, A., Paschali, M., Kumar, A., Johnston, A., Langlotz, C., Daneshjou, R., Koyejo, S., Alsentzer, E., and Chaudhari, A. S. MedVAL: Toward expert-level medical text validation with language models. *arXiv preprint arXiv:2507.03152*, 2025.

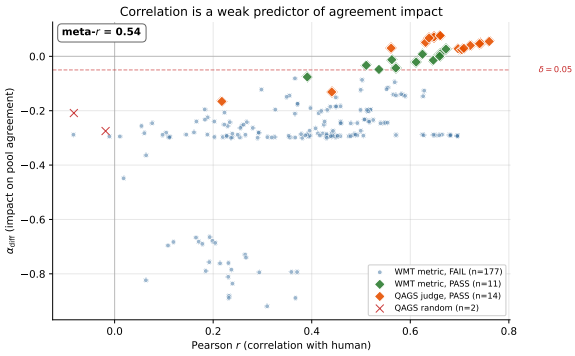

Figure 4: Each point is one metric or judge (188 WMT MQM + 16 QAGS). Diamonds pass α non-inferiority; blue circles are failing WMT metrics. Meta-correlation between Pearson r and α_{diff} is only 0.54; QAGS LLM judges (orange diamonds) improve human α despite only moderate correlation ($r \in [0.56, 0.76]$); random baseline (red $\times$) lies far below.

A Additional Figures

A.1 Pearson r vs. α_{diff} scatter

A.2 Task-adaptive margins by configuration

A.3 Domain-level and campaign-level summaries

A.4 199-Judge Threshold Analysis

We swept correlation thresholds from $r = -0.2$ to $r = 1.0$ across all 199 judge$\times$configuration rows (Figure 8). Per-domain optima differ substantially: QAGS is best at $r > 0.45$, SummEval at $r > 0.40$, WMT25 ESA at $r > 0.35$, and WMT MQM at $r > 0.50$. The optimal threshold spans a range of 0.15 across domains, and SummEval and WMT25 retain substantial error floors at *any* threshold. On SummEval specifically, a threshold that works well on one dimension fails on another: Opus degrades consistency by -0.138 and fails, while Flash degrades coherence by -0.147 and passes—because the human baseline and therefore the margin δ differ per dimension. The fundamental problem is that the threshold is domain-dependent, scale-dependent, and task-dependent—but practitioners must choose it *before* knowing these dependencies. The meta-correlation across all 199 configurations is $r_{\text{meta}} = 0.40$: correlation explains only 16% of the variance in agreement impact. Alpha Parity eliminates threshold selection entirely: each judge$\times$task configuration receives its own statistically grounded verdict, requiring no externally imposed cutoff.

Table 3: Task-adaptive non-inferiority margins $\delta = \max(2\sigma_{\text{LOO}}, 0.05)$ across representative configurations. The margin adapts to human rater variability: strict for tight expert pools (SummEval consistency, $\delta = 0.05$), lenient for noisy pools (WMT25 en-mas, $\delta = 0.55$).

Configuration	**Raters**	σ_{LOO}	δ	**Source**
QAGS CNN/DM	162	0.003	0.050	floor
QAGS XSum	84	0.008	0.050	floor
SummEval consistency	3	0.019	0.050	floor
SummEval fluency	3	0.031	0.063	computed
SummEval relevance	3	0.097	0.193	computed
SummEval coherence	3	0.105	0.210	computed
WMT23 en-de (MQM)	10	0.017	0.050	floor
WMT25 cs-de	12	0.012	0.050	floor
WMT25 cs-uk	7	0.060	0.120	computed
WMT25 en-mas	4	0.274	0.548	computed

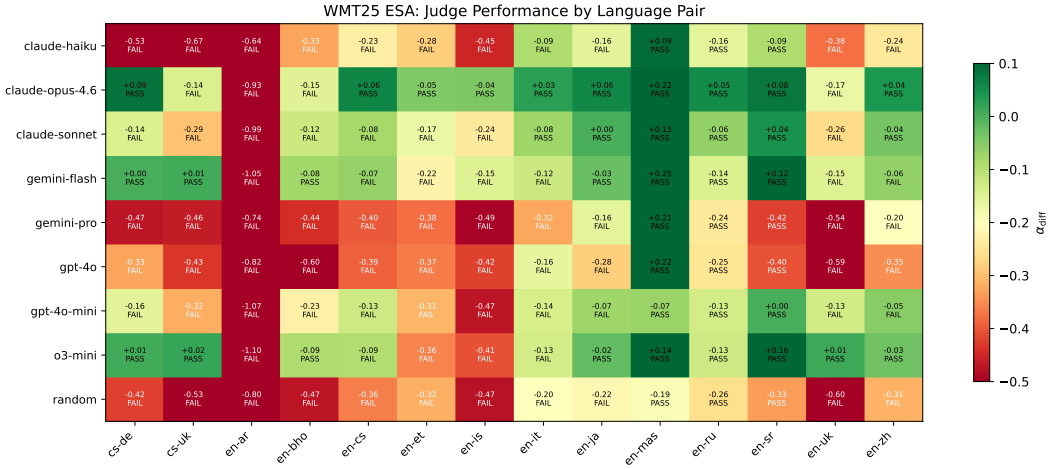

Figure 5: Full α_{diff} matrix for the WMT25 ESA judge evaluation. Rows are judges, columns are language pairs. Outlined cells indicate a PASS at the task-adaptive margin δ.

B Full Agreement Landscape Tables

B.1 WMT Direct Assessment (WMT21–WMT24)

Table 4 lists the 31 WMT academic DA cells drawn from the public WMT-DA legacy mirror, a subset of the 56-cell DA pool reported in §4.1 (the remaining 25 DA cells come from ToShip 2020 Microsoft commercial DA, summarized in Appendix B.4). α_{raw} uses raw DA scores; α_{znorm} applies z-normalization per rater to remove scale bias. σ_{LOO} is the leave-one-out standard deviation of α that calibrates the non-inferiority margin δ.

B.2 SummEval (Fabbri et al. 2021)

Table 5 shows agreement across the four SummEval quality dimensions for both expert annotators (trained NLP researchers, $n = 3$) and crowdworkers ($n = 5$). Expert consistency ($\alpha = 0.90$ interval) is the sole configuration in the entire landscape that clearly exceeds Krippendorff's 0.667 threshold.

B.3 QAGS (Wang et al. 2020)

Table 6 shows QAGS inter-rater agreement at both the summary level (interval: fraction of consistent sentences) and the sentence level (nominal: binary consistent/inconsistent). CNN/DM has substantially higher agreement than XSum, reflecting the harder factuality judgments on more abstractive summaries.

B.4 ToShip21 Machine Translation (Kocmi et al. 2021)

Table 7 reports summary statistics for the 55 language pairs in the ToShip21 meta-evaluation dataset (Kocmi et al. 2021) with at least one measured campaign. For brevity we show the 10 highest-agreement and 10 lowest-agreement LPs; the complete 55-LP table is available in the released data.

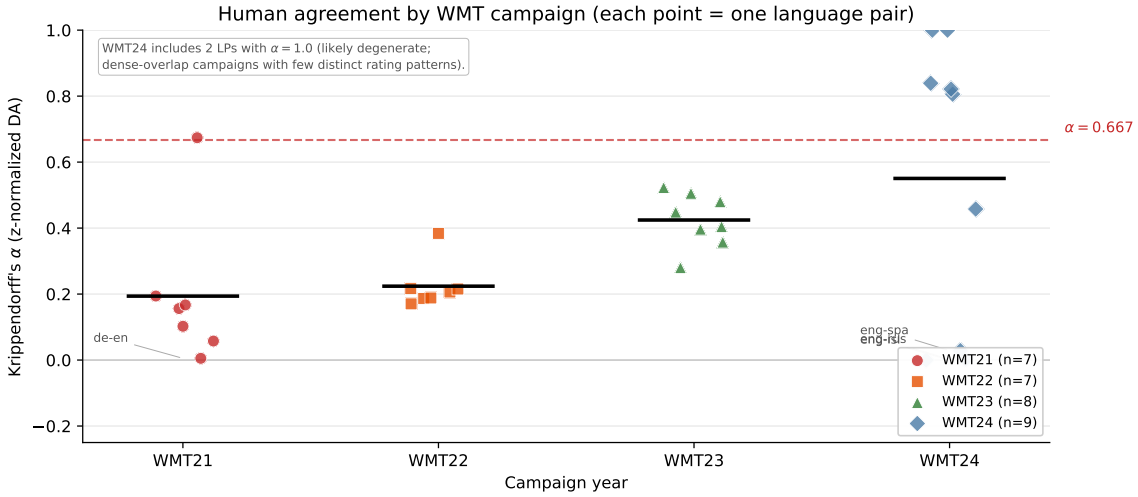

Figure 6: WMT DA z-normalized α by campaign year, one point per language pair. Black bars mark within-year means. WMT23 is the highest-quality DA campaign in our landscape (mean $\alpha \approx 0.42$).

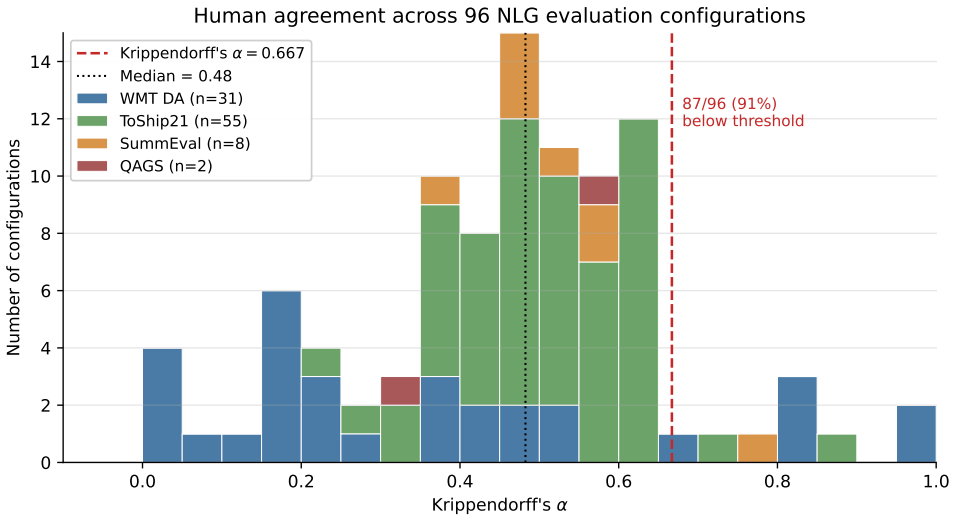

Figure 7: Distribution of human α across all 96 evaluation configurations. Bulk concentrates between 0.35 and 0.55; 87/96 (91%) fall below the $\alpha = 0.667$ threshold.

A cautionary cell at scale: `2022 fr-en DA`. The same corpus contains `2022 fr-en DA`: 66,536 units, 390 annotators, 14.6% overlap, $\alpha = 0.27$. This is the Microsoft commercial pool operating at full scale, and the agreement is weak—scale alone did not raise α, nor did a trained professional annotator pool with internal calibration, nor did the DA protocol itself on its own terms. Free scalar scoring on translation quality retains more inter-annotator noise than rubric-constrained protocols, even at industrial volume. This cell anchors the need for per-cell non-inferiority margins rather than a global calibration: if the candidate judge's α must stay within 0.05 of a 0.42 pool mean, the margin is comfortably resolvable, but within 0.05 of a 0.27 cell mean, the signal budget is much tighter and δ would need to condition on the stratum rather than the domain.

C Human Parity Retrospective

In 2018, Hassan et al. (2018) reported machine translation had achieved human parity on Chinese-to-English news, based on a significance test: mean DA scores for MT outputs did not differ significantly from those for human references. We revisit this claim through Alpha Parity. The combined α across all six rounds is 0.445 (z-normalized)—moderate but well below 0.667; a random anchor, constructed by shuffling each rater's score vector independently, achieves $\alpha = 0.251$ rather than zero because z-normalization preserves each rater's marginal distribution across the shuffle, so baseline-

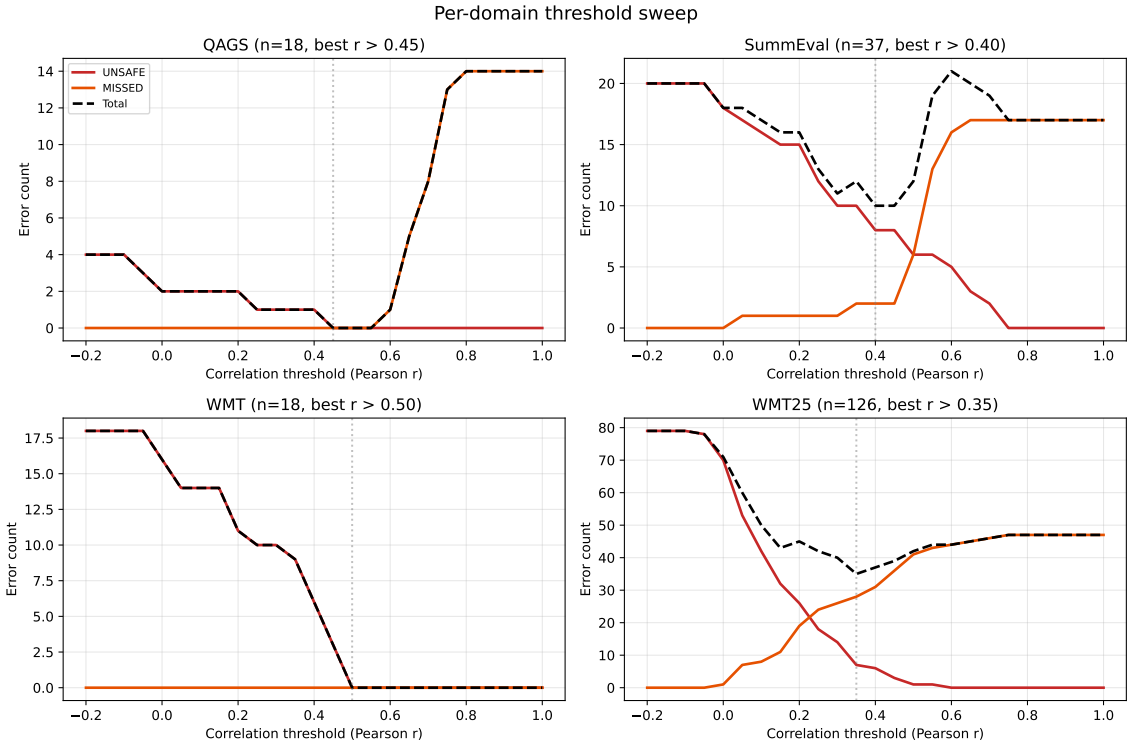

Figure 8: Per-domain threshold sweep across 199 judge×configuration rows. Each panel shows UNSAFE (deploy degraded judge), MISSED (reject deployable judge), and total errors as a function of the Pearson r threshold. The optimal threshold differs by domain—QAGS $r > 0.45$, SummEval $r > 0.40$, WMT MQM $r > 0.50$, WMT25 ESA $r > 0.35$—and SummEval and WMT25 retain substantial error floors even at their best threshold. No single correlation cutoff works across domains.

rate agreement is structurally retained even under random permutation. The effective agreement attributable to genuine consensus is therefore only 0.194 (LOO std $= 0.006$); the Stage 3 discriminative validity check (§3) operationalizes exactly this intuition—a random baseline provides the null of "agreement explainable by score distribution alone." Framed as non-inferiority ("is human-MT agreement not worse than human-human?"), the question would have been more informative: the significance test conflates two scenarios—(a) MT is as good as human translation, and (b) human raters cannot tell the difference because they don't agree much with each other (Kocmi et al., 2024b). This is not a criticism of the original study—which followed the methodological standards of its time—but illustrates how agreement-based analysis provides a richer foundation for evaluative claims.

D Beyond MT: Implications for Other Evaluation Settings

The validation gap generalizes beyond the domains we tested. In RLHF, reward models trained on LLM-generated preference labels inherit the judge's biases; Alpha Parity applied to preference datasets would quantify whether the labels are trustworthy training data, preventing the hill-climbing-downward failure mode flagged in §7. In safety evaluation, per-category α would reveal whether specific harm categories are systematically undetected—a uniform "LLM judge agrees 80% with human" headline can hide per-category agreement near zero on the categories that matter most. As AI regulation matures, organizations will increasingly need auditable evidence that evaluation processes are reliable. A non-inferiority test with a pre-declared margin provides exactly this: the null hypothesis is stated in advance, the bootstrap CI is reproducible from the annotated data, and the verdict is binary and defensible. We anticipate these non-MT applications to motivate domain-specific extensions of Alpha Parity (e.g., calibrating δ for preference data, or integrating Stage 3 checks with targeted red-teaming) rather than direct transplantation.

E Proof of Proposition 1

We restate the setting. Let $h_1, \ldots, h_n$ be a sample of human scores with finite mean μ_h and variance

Table 4: Human inter-rater agreement across 31 WMT academic DA campaign–LP cells (legacy public mirror).

Year	LP	Raters	Items	α_{raw}	α_{znorm}	σ_{LOO}
2021	cs-en	92	506	0.139	0.194	0.018
	de-en	182	1186	0.100	0.005	0.009
	ha-en	144	1320	0.100	0.156	0.013
	is-en	107	425	-0.056	0.674	0.015
	ja-en	133	1174	-0.526	0.058	0.041
	ru-en	96	1050	-0.565	0.167	0.041
	zh-en	235	3032	-0.034	0.102	0.017
2022	cs-en	156	1515	-0.074	0.170	0.025
	de-en	152	2356	-0.174	0.187	0.010
	ja-en	202	3723	0.321	0.384	0.009
	lv-en	46	2070	0.188	0.206	0.012
	ru-en	217	2112	0.113	0.215	0.008
	uk-en	144	7335	0.113	0.188	0.013
	zh-en	172	2836	0.052	0.217	0.009
2023	ces-ukr	146	15750	0.218	0.280	0.004
	deu-eng	82	10024	0.299	0.480	0.006
	eng-ces	162	19808	0.229	0.448	0.006
	eng-deu	87	9633	0.277	0.505	0.004
	eng-jpn	164	21046	0.374	0.405	0.005
	eng-zho	154	19808	0.600	0.523	0.004
	jpn-eng	174	23436	0.344	0.397	0.005
	zho-eng	128	16240	0.300	0.357	0.007
2024	ces-ukr	299	24524	1.000	1.000	0.000
	eng-ces	187	10730	0.825	0.806	0.002
	eng-hin	131	7474	0.530	0.458	0.005
	eng-isl	89	7304	0.000	0.000	0.000
	eng-jpn	212	9109	0.809	0.821	0.001
	eng-rus	114	9354	-0.001	0.002	0.000
	eng-spa	114	9354	-0.002	0.029	0.000
	eng-ukr	89	7304	1.000	1.000	0.000
	eng-zho	212	9109	0.815	0.839	0.001

$\sigma_h^2 > 0$, and let judge scores be generated as

$$j_i = a\,h_i + b + \epsilon_i, \quad a > 0,\ b \in \mathbb{R},\ \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

with $\{\epsilon_i\}$ independent of $\{h_i\}$. All expectations below are population-level (equivalently, large-n limits of sample moments). Without loss of generality we take $\mu_h = 0$; affine invariance of both coefficients under shifts of h lets us absorb any nonzero mean into b.

Part 1: Pearson $\to 1$ as $\sigma \to 0$. From the definitions,

$$\begin{aligned} \mathrm{Cov}(h, j) &= \mathrm{Cov}(h, ah + b + \epsilon) = a\,\sigma_h^2, \\ \mathrm{Var}(j) &= a^2\sigma_h^2 + \sigma^2, \end{aligned}$$

using independence of ϵ from h. The Pearson correlation is therefore

$$\begin{aligned} r(h, j) &= \frac{\mathrm{Cov}(h, j)}{\sigma_h \sqrt{\mathrm{Var}(j)}} = \frac{a\,\sigma_h^2}{\sigma_h \sqrt{a^2\sigma_h^2 + \sigma^2}} \\ &= \frac{a\,\sigma_h}{\sqrt{a^2\sigma_h^2 + \sigma^2}}. \end{aligned} \tag{2}$$

Because $a > 0$, taking $\sigma \to 0$ in (2) yields

$$\lim_{\sigma \to 0} r(h, j) \;=\; \frac{a\sigma_h}{a\sigma_h} \;=\; 1,$$

and this limit is independent of a and b: b does not appear in (2) at all (Pearson is translation-invariant), and the positive scaling factor a cancels between numerator and denominator.

Part 2: Krippendorff's α at interval level strictly decreases with $|a - 1|$ or $|b|$. Krippendorff's α with two raters and interval-level distance function $d(x, y) = (x - y)^2$ admits the closed form (Krippendorff, 2011)

$$\alpha \;=\; \frac{2\,\mathrm{Cov}(h, j)}{\mathrm{Var}(h) + \mathrm{Var}(j) + (\mu_h - \mu_j)^2}. \tag{3}$$

This is a population-level statement that coincides with the sample α in the large-n limit, and it equals Lin's concordance correlation coefficient (Lin, 1989). Substituting the moments computed in Part 1 and using $\mu_h = 0$, $\mu_j = b$:

$$\alpha(a, b, \sigma) \;=\; \frac{2a\sigma_h^2}{(1 + a^2)\sigma_h^2 + b^2 + \sigma^2}. \tag{4}$$

Dependence on b. Fixing $a > 0$ and $\sigma \geq 0$, let $f(b) = \alpha(a, b, \sigma)$. Differentiating (4) with respect to b^2,

$$\frac{\partial \alpha}{\partial (b^2)} \;=\; -\frac{2a\sigma_h^2}{\left[(1 + a^2)\sigma_h^2 + b^2 + \sigma^2\right]^2} \;<\; 0,$$

so α is a strictly decreasing function of b^2, and therefore of $|b|$, for every fixed $a > 0$.

Dependence on a. Fixing b and σ, let $g(a) = \alpha(a, b, \sigma)$, $c = \sigma_h^2$, $k = b^2 + \sigma^2$. Then

$$g(a) \;=\; \frac{2ac}{(1 + a^2)c + k}.$$

Differentiating,

$$\begin{aligned} g'(a) &= \frac{2c\left[(1 + a^2)c + k\right] - 2ac \cdot 2ac}{\left[(1 + a^2)c + k\right]^2} \\ &= \frac{2c\left[(1 - a^2)c + k\right]}{\left[(1 + a^2)c + k\right]^2}. \end{aligned}$$

Setting $g'(a) = 0$ gives $(1 - a^2)c + k = 0$, i.e., $a^2 = 1 + k/c$, so the unique positive critical point is $a^* = \sqrt{1 + k/c} \geq 1$, with equality iff $k = 0$ (no offset, no noise). One can verify $g''(a^*) < 0$, so a^* is a maximum. In the ideal case $b = 0$, $\sigma = 0$ (i.e. $k = 0$), we have $a^* = 1$ and $g(1) = 1$. For any fixed $k > 0$, g is strictly unimodal with its maximum at $a^* > 1$; in particular g is

Table 5: SummEval inter-rater agreement by annotator type and dimension.

Rater type	Dimension	Raters	Items	α_{ordinal}	α_{interval}	σ_{LOO}
Expert	coherence	3	1600	0.554	0.559	0.105
	consistency	3	1600	0.796	0.899	0.018
	fluency	3	1600	0.588	0.726	0.031
	relevance	3	1600	0.397	0.453	0.097
Crowd	coherence	5	1600	0.499	0.494	0.006
	consistency	5	1600	0.506	0.495	0.006
	fluency	5	1600	0.492	0.483	0.005
	relevance	5	1600	0.485	0.478	0.005

Table 6: QAGS inter-rater agreement by subset and analysis level.

Subset	Level	Scale	Workers	Items	Raters/item	α	Prev.
CNNDM	summary	interval	162	235	3.0	0.567	0.451
	summary_binary	nominal	162	235	3.0	0.468	0.451
	sentence	nominal	162	714	3.0	0.513	0.720
XSUM	summary	interval	84	239	3.0	0.342	0.485
	summary_binary	nominal	84	239	3.0	0.342	0.485
	sentence	nominal	84	239	3.0	0.342	0.485

strictly increasing on $(0, a^*)$ and strictly decreasing on (a^*, ∞). Evaluating at $a = 1$,

$$g(1) \;=\; \frac{2c}{2c+k} \;<\; 1 \quad \text{whenever } k > 0,$$

and the gap to 1 grows in k. Thus for any realistic perturbation—even infinitesimal offset $b \neq 0$ or noise $\sigma > 0$—the value of α is bounded strictly below the Pearson limit of 1, and for noise-free scale deviations ($b = 0$, $\sigma = 0$, $a \neq 1$), $g(a) < g(1) = 1$ strictly as $|a - 1|$ increases.

Consequence. Combining Parts 1 and 2, the same affine transformation family $j_i = ah_i + b + \epsilon_i$ that drives Pearson correlation to 1 (as noise vanishes) drives Krippendorff's α strictly below 1. The two coefficients answer different questions: Pearson asks whether the judge's scores are a positive linear function of the human's; α asks whether the judge's scores *agree with* the human's on an absolute scale. Any non-identity affine transformation is invisible to the former and detectable by the latter. This formalizes the intuition underlying the adversarial Bias A in Appendix H (scale compression by $\lambda = 0.3$): Pearson r is provably unchanged, while α drops and the Alpha Parity verdict correctly flips from PASS to FAIL. □

F Bootstrap Procedure for Non-Inferiority Testing

This appendix describes the bootstrap procedure used to test whether adding an LLM judge to the human rater pool degrades Krippendorff's α beyond a pre-registered margin δ. The procedure adapts established bootstrap methods for agreement coefficients (Krippendorff, 2011; Hayes and Krippendorff, 2007) to the non-inferiority testing framework from clinical trials (Piaggio et al., 2012; Wellek, 2010).

F.1 Setup

Let $\mathbf{M} \in \mathbb{R}^{R \times N}$ be the reliability matrix, where R is the number of human raters and N is the number of items (e.g., system–segment pairs). Entry $m_{r,i}$ is the score assigned by rater r to item i, with $m_{r,i} = \texttt{NaN}$ for missing values (rater r did not rate item i). Let $\mathbf{j} \in \mathbb{R}^N$ be the judge's score vector, with j_i the score assigned to item i.

Define the augmented matrix $\mathbf{M}^+ \in \mathbb{R}^{(R+1) \times N}$ that adds the judge as row $R+1$:

$$\mathbf{M}^+ = \begin{bmatrix} \mathbf{M} \\ \mathbf{j}^\top \end{bmatrix}. \tag{5}$$

The quantities of interest are the human-only pool agreement $\alpha_H = \alpha(\mathbf{M})$, the agreement with the judge included $\alpha_{H+J} = \alpha(\mathbf{M}^+)$, and the agreement difference $d = \alpha_{H+J} - \alpha_H$.

The null and alternative hypotheses for the non-inferiority test are:

$$H_0 : d < -\delta \quad \text{(judge is inferior by more than } \delta) \tag{6}$$

$$H_1 : d \geq -\delta \quad \text{(judge is non-inferior)} \tag{7}$$

Table 7: Top 10 and bottom 10 language pairs by mean z-normalized α across ToShip21 campaigns.

LP	Campaigns	Raters (mean)	Overlap (mean)	α_{raw}	α_{znorm}
Top 10					
MLT-ENU	1	5.0	220	0.850	0.851
MLG-ENU	1	5.0	220	0.581	0.728
ENU-RUS	11	9.5	248	0.500	0.647
ENU-MLG	1	5.0	221	0.592	0.645
ENU-KOR	11	9.5	232	0.472	0.636
SMO-ENU	1	5.0	222	0.640	0.628
ENU-TRK	10	10.0	212	0.364	0.623
KOR-ENU	10	10.0	220	0.469	0.621
ENU-CSY	10	10.0	246	0.332	0.608
CSY-ENU	10	10.0	230	0.325	0.608
Bottom 10					
ESN-ENU	10	10.0	250	0.158	0.397
PTB-ENU	10	10.0	250	0.161	0.381
IND-ENU	10	10.0	236	0.175	0.379
ENU-FRA	10	10.0	250	0.177	0.373
ENU-SMO	1	5.0	225	0.326	0.370
ENU-HIN	11	9.5	247	0.136	0.361
ITA-ENU	10	10.0	225	0.149	0.334
FRA-ENU	10	10.0	250	0.258	0.308
ENU-ITA	10	10.0	250	0.113	0.295
CHS-YUE	1	10.0	225	-0.020	0.224

F.2 Bootstrap Algorithm

1. **Pre-register the margin δ.** We recommend $\delta = \max(2 \times \sigma_{\text{LOO}},\ 0.05)$, where σ_{LOO} is the leave-one-out standard deviation of α_H (see Section F.3 below). This makes δ task-adaptive: stricter for high-agreement configurations, more lenient for noisy ones.

2. **Generate B bootstrap samples** (we use $B =$ 10,000). For each $b = 1, \ldots, B$:

   (a) Sample N item indices with replacement: $\mathcal{I}^{(b)} = \{i_1^{(b)}, \ldots, i_N^{(b)}\}$, where each $i_k^{(b)} \sim \text{Uniform}(1, N)$.

   (b) Extract resampled matrices:

   $$\mathbf{M}^{(b)} = \mathbf{M}[:, \mathcal{I}^{(b)}] \quad (8)$$

   $$\mathbf{M}^{+,(b)} = \mathbf{M}^{+}[:, \mathcal{I}^{(b)}] \quad (9)$$

   (c) Compute agreement on resampled data:

   $$\alpha_H^{(b)} = \alpha(\mathbf{M}^{(b)}) \quad (10)$$

   $$\alpha_{H+J}^{(b)} = \alpha(\mathbf{M}^{+,(b)}) \quad (11)$$

   (d) Record the difference: $d^{(b)} = \alpha_{H+J}^{(b)} - \alpha_H^{(b)}$.

3. **Construct the one-sided confidence interval.** Sort the differences $\{d^{(1)}, \ldots, d^{(B)}\}$. The γ-level lower bound of the one-sided CI is:

   $$\text{CI}_{\text{lower}}^{\gamma} = d_{(\lfloor (1-\gamma) \cdot B \rfloor)} \quad (12)$$

   where $d_{(k)}$ is the k-th order statistic and $\gamma = 0.95$ for a 95% one-sided CI.

4. **Decision.** The judge passes non-inferiority if:

   $$\text{CI}_{\text{lower}}^{0.95} > -\delta \quad (13)$$

Critical design choices.

- **Resample items, not raters.** The bootstrap resamples columns (items), not rows (raters). This preserves the within-item correlation structure: if two raters agree on item i, that agreement is maintained in the bootstrap sample. Resampling raters would break the paired structure of the reliability matrix.

- **Block bootstrap for document structure.** When items have document-level grouping (e.g., multiple segments from the same article in SummEval, or multiple system outputs for the same source in WMT), resample at the document level rather than the item level. This accounts for within-document correlation. Concretely: sample D document indices with replacement, then include all items from each selected document.

- **Paired differences.** The quantity $d^{(b)}$ is computed on the *same* bootstrap sample for both α_H and α_{H+J}. This pairing reduces variance compared to computing each independently.

- **Handling degenerate bootstraps.** Some bootstrap samples may produce undefined α (e.g., if all resampled items have the same score). Discard these and note the discard rate. If more than 5% of bootstraps are degenerate, the dataset may be too small for reliable inference.

F.3 Leave-One-Out Calibration of δ

The margin δ should reflect natural variation in human agreement. We calibrate it using the leave-one-out (LOO) standard deviation:

1. For each human rater $r = 1, \ldots, R$:
   (a) Remove row r from $\mathbf{M}$ to form $\mathbf{M}^{(-r)} \in \mathbb{R}^{(R-1)\times N}$.
   (b) Compute $\alpha^{(-r)} = \alpha(\mathbf{M}^{(-r)})$.
2. Compute $\sigma_{\text{LOO}} = \text{std}(\alpha^{(-1)}, \ldots, \alpha^{(-R)})$.
3. Set $\delta = \max(2 \times \sigma_{\text{LOO}},\ 0.05)$.

Interpretation. σ_{LOO} measures how much α fluctuates when a single human rater is removed. A judge that degrades α by less than $2 \times \sigma_{\text{LOO}}$ is indistinguishable from the natural variation caused by replacing one human rater with another. The floor of 0.05 ensures that the test has non-trivial power even when LOO variation is very small (e.g., when all raters are near-identical).

Example from data. On WMT23 en-de (MQM, 10 raters), $\sigma_{\text{LOO}} = 0.006$, yielding $\delta = \max(0.012, 0.05) = 0.05$. On QAGS CNNDM (162 crowd workers), $\sigma_{\text{LOO}} = 0.003$, yielding $\delta = 0.05$. The floor dominates in both cases because individual raters contribute little marginal variance when the pool is large.

F.4 Sensitivity Analysis

We recommend reporting results at three margins:

Margin	Interpretation
$\delta/2$	Strict: judge close to human agreement
δ	Primary: data-calibrated default
2δ	Lenient: allows larger degradation

A judge that passes at $\delta/2$ is a strong substitute. A judge that passes at 2δ but fails at δ is marginal. A judge that fails at 2δ is clearly inferior. This three-tier reporting avoids dependence on a single δ choice.

F.5 Power Considerations

The test has sufficient power when:

1. The number of items N is large enough. With $N < 100$, bootstrap CIs are wide and the test may fail to reject H_0 even for a good judge. We recommend $N \geq 500$ for segment-level evaluation.
2. The number of raters R is at least 3 (ideally 5+). With $R = 2$, α_H is estimated from a single pairwise comparison, which is inherently unstable.
3. δ is not too small. Very tight margins (e.g., $\delta = 0.01$) require very large N to achieve power. The LOO-calibrated δ avoids this by scaling to the data.

Random anchor as power check. The distribution-matched random annotator must fail the non-inferiority test. If it passes, the test lacks power—either N is too small, δ is too large, or the human baseline α_H is itself too low for meaningful discrimination. We report random anchor results for every configuration.

F.6 Reference Implementation

G Complete Pass/Fail Matrix

Tables in this appendix list α_{diff} for every judge×configuration combination in our experiments. **Bold values** indicate a PASS verdict at the task-adaptive non-inferiority margin δ; plain values fail. Entries shown as "–" indicate the judge was not run on that configuration.

G.1 QAGS (Factual Consistency)

Table 8: QAGS pass/fail matrix: α_{diff} with binary nominal α for factual consistency.

Judge	cnndm	xsum
Haiku	**+0.049**	**+0.071**
Sonnet	**+0.046**	**+0.071**
Opus4.6	**+0.028**	**+0.076**
Flash	**+0.025**	**+0.030**
Pro	-0.131	-0.166
4o-mini	**+0.040**	**+0.068**
GPT-4o	**+0.055**	**+0.051**
o3-mini	**+0.029**	**+0.076**
Random	-0.275	-0.209

Table 9: SummEval pass/fail matrix: α_{diff} at ordinal scale for four quality dimensions.

Judge	coherence	consistency	fluency	relevance
Haiku	**-0.047**	-0.200	-0.527	**-0.021**
Sonnet	**-0.049**	-0.200	-0.576	**-0.074**
Opus4.6	**-0.029**	-0.138	-0.474	**+0.000**
Flash	**-0.147**	-0.132	-0.498	**-0.032**
Pro	-0.509	-0.788	-0.621	**-0.055**
4o-mini	**-0.163**	-0.511	-0.555	**-0.051**
GPT-4o	**-0.174**	-0.495	-0.509	**-0.177**
o3-mini	**-0.047**	-0.129	-0.541	**-0.049**
Random	-0.263	-0.539	-0.395	**-0.190**

Table 10: WMT23 MQM pass/fail matrix. Construct mismatch: all judges fail.

Judge	wmt23_en-de	wmt23_zh-en
Haiku	-0.855	-0.443
Sonnet	-0.873	–
Opus4.6	-0.835	–
Flash	-0.837	–
Pro	-0.817	–
4o-mini	-0.876	–
GPT-4o	-0.715	–
o3-mini	-0.886	–
Random	-0.818	–

G.2 SummEval (Summarization Quality)

G.3 WMT23 MQM (Machine Translation, Expert Annotation)

G.4 WMT25 ESA (Machine Translation, Error Span Annotation)

The WMT25 ESA matrix is split across two tables for readability.

Table 11: WMT25 ESA pass/fail matrix, LPs 1–6.

Judge	cs-de	cs-uk	en-ar	en-bho	en-cs	en-et
Haiku	-0.525	-0.671	-0.637	-0.327	-0.232	-0.280
Sonnet	-0.142	-0.300	-0.985	-0.123	-0.082	-0.170
Opus4.6	**+0.086**	-0.142	-0.925	-0.151	**+0.062**	**-0.046**
Flash	**+0.003**	**+0.006**	-1.049	**-0.080**	-0.066	-0.220
Pro	-0.472	-0.464	-0.737	-0.442	-0.399	-0.381
4o-mini	-0.157	-0.321	–	–	–	–
GPT-4o	-0.326	-0.432	–	–	–	–
o3-mini	**+0.008**	**+0.019**	–	–	–	–
Random	-0.423	-0.531	-0.800	-0.470	-0.362	-0.317

Table 12: WMT25 ESA pass/fail matrix, LPs 7–12.

Judge	en-is	en-it	en-ja	en-mas	en-ru	en-sr
Haiku	-0.452	-0.091	–	–	–	–
Sonnet	-0.239	-0.077	–	–	–	–
Opus4.6	**-0.043**	**+0.027**	–	–	–	–
Flash	-0.150	-0.123	**-0.025**	**+0.239**	**-0.145**	**+0.125**
Pro	-0.495	-0.325	-0.158	**+0.212**	**-0.242**	**-0.423**
Random	-0.474	-0.198	-0.215	–	–	–

Table 13: Adversarial bias injection. GPT-4o's actual QAGS CNN/DM scores are distorted with three realistic biases. Bias A preserves r exactly (Proposition 1) while degrading α. All biased variants fail Alpha Parity.

Variant	r	α_{diff}	Verdict
GPT-4o (actual)	0.760	+0.055	PASS
A: Scale compression ($\times 0.3$)	0.760	-0.062	FAIL
B: Positional flip (40%)	0.564	-0.036	FAIL
C: Conditional flip (30%)	0.445	-0.101	FAIL
D: All combined	-0.090	-0.258	FAIL

H Adversarial Bias Injection Details

This appendix formalizes the four bias-injection procedures summarized in §5.1. The setup is fixed throughout: let $J = (j_1, \ldots, j_n)$ denote GPT-4o's actual binary factuality predictions on the QAGS CNN/DM subset ($n = 714$ sentences), and let $Y = (y_1, \ldots, y_n)$ denote the corresponding majority-vote human labels. The baseline configuration—no injection—achieves $r = 0.760$ and $\alpha_{\text{diff}} = +0.055$ (PASS).

A: Scale compression ($\lambda = 0.3$). Each binary score $j_i \in \{0, 1\}$ is mapped to the real value $j'_i = \lambda \cdot j_i + (1-\lambda) \cdot 0.5$, yielding $j'_i \in \{0.35, 0.65\}$. The transformation is affine with $\lambda > 0$, so by Proposition 1, Pearson r and Spearman ρ are invariant under this mapping: the ranking of items is preserved and the linear relationship between J and Y has identical slope structure. However, Krippendorff's α at interval level responds to absolute distances: the effective inter-rater disagreement is now compressed toward the midpoint, artificially inflating observed agreement on items where raters disagree and deflating it where they agree. The net effect on α_{diff} is a drop of 0.117 (from $+0.055$ to -0.062), flipping the verdict from PASS to FAIL while correlation-based validation records no change.

B: Positional flip (40%). Let $I \subset \{1, \ldots, n\}$ be a deterministic set of indices of size $\lceil 0.40n \rceil = 286$, constructed by taking every $\lceil n/286 \rceil$-th position so that flipped and unflipped cases are evenly distributed across the dataset. For $i \in I$, set $j''_i = 1 - j_i$; otherwise $j''_i = j_i$. The flips are independent of content—a "content-blind attacker" who degrades the judge uniformly. Both r and α degrade in step: r drops to 0.564 and α_{diff} to -0.036. Either approach would correctly reject this variant at common thresholds.

C: Conditional flip (30%). This injection targets items where the judge and the majority-vote label already disagree: $I = \{i : j_i \neq y_i\}$, from which we flip a random 30% subsample $I' \subseteq I$.

The result is a judge that is *systematically* worse at the examples where it was already wrong—a content-sensitive failure mode that can emerge in fine-tuned judges overfit to easy cases. Pearson r drops to 0.445 while α_{diff} drops to -0.101, and importantly the α degradation is larger than B's despite a smaller fraction flipped, because the flips concentrate in disagreement regions that matter more for pairwise agreement.

D: Combined (A $\circ$ B $\circ$ C). Apply scale compression with $\lambda = 0.3$, followed by positional flip at 40%, followed by conditional flip at 30%. The combined transformation destroys both ranking and agreement structure: $r = -0.090$ and $\alpha_{\text{diff}} = -0.258$. This is the "fully degraded" baseline against which the other injections are calibrated.

Seed and reproducibility. All stochastic injections (B, C, D) use a fixed random seed (`seed=42`). The injection script will be released alongside the toolkit upon acceptance; running it against GPT-4o's released QAGS predictions reproduces Table 13 exactly.

Interpretation. The adversarial pattern isolates where correlation-based validation fails in a way that is not a simple artifact of noisy data: injection A degrades the judge's measurement scale without altering any rank or linear trend, exactly the kind of manipulation that correlation is structurally blind to. That Alpha Parity detects the manipulation (verdict flips from PASS to FAIL) while r does not reflects the fundamental methodological claim of this paper: correlation answers a different question from agreement, and only the agreement-framed question can catch scale-based bias.

I Extended Related Work

Correlation critique in MT evaluation. Kocmi et al. (Kocmi et al., 2021) showed pairwise accuracy outperforms Pearson/Spearman correlation for system-level deployment on 2.3M sentence-level judgments; Kocmi et al. (Kocmi et al., 2024b) mapped metric-score deltas to pairwise-accuracy thresholds and showed p-values become trivial with large test sets; Deutsch et al. (Deutsch et al., 2023) refined segment-level meta-evaluation with tie-aware pairwise accuracy, replacing Kendall's τ; DiIanni and Deutsch (DiIanni and Deutsch, 2025) propose Pairwise-Difference Pearson (PDP), refining Global Pearson to intra-segment pairwise differences. Moghe et al. (Moghe et al., 2023) extend the correlation critique to extrinsic validity: segment-level metric–human agreement shows negligible correlation with downstream task success. Perrella et al. (Perrella et al., 2024a) diagnose structural issues in the WMT Metrics Shared Task meta-evaluation framework itself via "sentinel" metrics, and reframe MT-metric evaluation around interpretable Precision/Recall/F-score applied to data-filtering and re-ranking scenarios (Perrella et al., 2024b); most recently, Proietti et al. (Proietti et al., 2025) incorporate human baselines into WMT meta-evaluation and find that state-of-the-art metrics match or exceed humans, questioning whether MT-eval progress can still be measured by correlation-with-human at all. Flamich et al. (Flamich et al., 2025) prove information-theoretically that any single-score MT evaluation cannot jointly capture accuracy and naturalness, further motivating multi-dimensional validation. Complementary work documents specific LLM-judge failure modes invisible to correlation-based validation (error-type sensitivity (Moghe et al., 2024), MBR metric bias (Kovacs et al., 2024), length-dependent degradation (Zouhar et al., 2025), poor segment-level scoring (Cui, 2025), LRM overthinking and scoring overestimation (Zhan et al., 2025), single-hub adversarial texts beating M2M100 on COMET (Deguchi et al., 2026)) and proposes alternative protocols for judge construction (error-span prompting (Fernandes et al., 2023), preference reward modeling (Tan and Monz, 2025), test-set specialization (Finkelstein et al., 2024), knockout tournaments (Sandan et al., 2025), synthetic annotations (Wang et al., 2026), synthetic meta-correlation (Eigler et al., 2026), multilingual judge meta-evaluation (Son et al., 2024)); Alpha Parity is agnostic to these choices. Our complementary gap is *segment-level substitutability*: system-level ranking accuracy $\rightarrow$ delta-accuracy $\rightarrow$ tie-aware pairwise accuracy $\rightarrow$ rater substitutability (this work).

Concurrent critiques of correlation for judge validation. Han et al. (Han et al., 2025) pair correlation filtering ($r \geq 0.80$) with Cohen's κ and a human-likeness z-score on 54 LLMs; Song et al. (Song et al., 2026) show model-level $\rho = 0.99$ masks sample-level ICC $= 0.67$; Landesberg (Landesberg, 2026) shows global $r = 0.47$ captures only 21% of best-of-N gain; Nasser et al. (Nasser et al., 2026) report $\alpha = 0.042$ across nine frontier

judges ("reliability paradox"); Choi et al. (Choi et al., 2026) apply IRT-GRM for diagnostic assessment. Alpha Parity differs by (1) deriving δ from human LOO variability rather than imposing a fixed correlation filter; (2) using Krippendorff's α for scale-adaptive agreement (Cohen's κ handles only categorical); (3) non-inferiority framing yields auditable pass/fail verdicts; (4) cross-domain evaluation revealing construct mismatch and quality-enhancer regimes.

Measurement science and clinical precedent. Bland and Altman (Bland and Altman, 1986, 1999) established correlation cannot assess agreement; Artstein and Poesio (Artstein and Poesio, 2008) surveyed chance-corrected measures for NLP; Krippendorff (Krippendorff, 2004, 2011) developed α. Rosso (Rosso, 2015) and Edwards et al. (Edwards et al., 2021) independently rediscovered the distinction in radiology and sonography but stopped at recommending better metrics. Yamagishi et al. (Yamagishi et al., 2026) evaluate LLM-judge assessments of LLM-generated Japanese translations of chest CT reports against blinded radiologists, finding near-zero QWK agreement (-0.04 to 0.15) and strong LLM-judge self-preference ($> 93\%$) for LLM output—a cross-domain instance of the structural bias correlation cannot detect. Two clinical works are most directly parallel: Grolleau et al. (Grolleau et al., 2025) (MedFactEval) validate an LLM Jury for clinical summaries via Cohen's κ with a formal non-inferiority test against a pre-specified margin, and Aali et al. (Aali et al., 2025) (MedVAL) train a distilled evaluator that achieves macro-F1 non-inferior to a single human expert on 6 medical tasks ($\Delta_{\text{LCB-95}} > -0.05$)—both among the few prior works combining chance-corrected agreement with a formal non-inferiority hypothesis test for LLM judges, and both confined to a single clinical task with binary or categorical labels. Alpha Parity generalizes this to a scale-adaptive three-stage framework across four NLP domains with task-adaptive δ and a discriminative validity check.

Broader LLM-judge validity critique and certification. Leiter and Eger (Leiter and Eger, 2025) argue from measurement theory that LLJ adoption outpaces validation; Schaeffer et al. (Schaeffer et al., 2026) advocate ICC over Pearson; Thakur et al. (Thakur et al., 2025) show correlation overestimates agreement for lacking chance correction. Shen et al. (Shen et al., 2026) propose constrained-MLE failure-rate certification as complement to PPI (Angelopoulos et al., 2023). Alpha Parity is upstream: they estimate failure rates given a biased judge; we gate which judges are safe.

The parity debate and evaluation reform. Hassan et al. (Hassan et al., 2018) claimed human parity on Chinese-English news; Läubli et al. (Läubli et al., 2018) challenged this via document-level evaluation. Subsequent reforms—MQM (Freitag et al., 2021) and ESA (Kocmi et al., 2024d)—improved annotation quality (WMT24 ESA α up to 0.84 vs. 0.42 for WMT23 DA). For literary translation specifically, Zhang et al. (Zhang et al., 2025) (LITEVAL-CORPUS) demonstrate that MQM itself is inadequate as a human gold standard, with student MQM annotators preferring human translations over the best LLM output in only $\sim$42% of cases (vs. $\sim$82% with BWS and $\sim$94% with professional SQM)—a reminder that the measurement instrument, not only the judge, can be mis-specified. Alpha Parity closes the loop by validating automated substitutes for improved raters.

LLM-as-a-Judge and biases. Prior work on LLM judges—MT-Bench (Zheng et al., 2023), Prometheus (Kim et al., 2024), GEMBA (Kocmi and Federmann, 2023), and surveys (Gu et al., 2024)—validates via correlation, not agreement-based pass/fail testing. Documented biases include positional effects (Wang et al., 2023), self-preference (Panickssery et al., 2024), and inconsistency (Stureborg et al., 2024)—precisely the biases correlation cannot detect (Proposition 1).

MT metrics and evaluation datasets. Learned metrics (Rei et al., 2020) dominate WMT shared tasks (Freitag et al., 2022; Freitag et al., 2023) ranked by correlation—the very practice we argue is insufficient. Freitag et al. (Freitag et al., 2021) motivated MQM. We use SummEval (Fabbri et al., 2021), QAGS (Wang et al., 2020), and the Hassan et al. (Hassan et al., 2018) bilingual DA dataset as cross-domain test beds.

```python
import numpy as np
import krippendorff

def bootstrap_noninferiority(
    matrix_with_judge, judge_row_index,
    level='interval', delta=0.10,
    n_bootstrap=10000, confidence=0.95,
    block_ids=None, seed=42
):
    rng = np.random.default_rng(seed)
    R_plus_1, N = matrix_with_judge.shape

    # Separate human and full matrices
    human_rows = np.delete(np.arange(R_plus_1), judge_row_index)
    M_human = matrix_with_judge[human_rows, :]
    M_full = matrix_with_judge

    # Determine resampling units
    if block_ids is not None:
        unique_blocks = np.unique(block_ids)
        n_blocks = len(unique_blocks)
        block_to_items = {b: np.where(block_ids == b)[0] for b in unique_blocks}

    differences = []
    n_degenerate = 0

    for b in range(n_bootstrap):
        # Resample
        if block_ids is not None:
            sampled_blocks = rng.choice(unique_blocks, size=n_blocks, replace=True)
            idx = np.concatenate([block_to_items[bl] for bl in sampled_blocks])
        else:
            idx = rng.choice(N, size=N, replace=True)

        M_h_b = M_human[:, idx]
        M_f_b = M_full[:, idx]

        try:
            a_h = krippendorff.alpha(reliability_data=M_h_b, level_of_measurement=level)
            a_f = krippendorff.alpha(reliability_data=M_f_b, level_of_measurement=level)
            differences.append(a_f - a_h)
        except:
            n_degenerate += 1

    differences = np.array(differences)
    gamma = 1 - confidence
    ci_lower = np.percentile(differences, 100 * gamma)

    return {
        'passed': ci_lower > -delta,
        'alpha_diff_mean': np.mean(differences),
        'ci_lower': ci_lower,
        'ci_upper': np.percentile(differences, 100 * (1 - gamma)),
        'delta': delta,
        'n_bootstrap': n_bootstrap,
        'n_degenerate': n_degenerate,
        'degenerate_rate': n_degenerate / (n_bootstrap + n_degenerate),
        'p_value': np.mean(differences < -delta),
    }
```

Figure 9: Reference implementation of the bootstrap non-inferiority procedure (Python, NumPy, `krippendorff` package).