QUANTIFYING CONSENSUS: OVERCOMING EXACT MATCH LIMITATIONS IN LLM-AS-A-JUDGE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As Large Language Models (LLMs) increasingly serve as autonomous evaluators for complex outputs (LLM-as-a-Judge), the demand for fine-grained evaluation on granular scales (e.g., 1-5 or 1-10) has grown. Yet, this paradigm shift has exposed a critical vulnerability: we are evaluating the alignment of the next-generation artificial judges using legacy metrics that were never designed for this task. Current evaluation methods primarily rely on rank correlation coefficients (e.g., Spearman's ρ) or traditional Inter-Annotator Agreement (IAA) metrics like Cohen's Kappa. Both approaches are fundamentally limited for modern LLM evaluation: rank metrics ignore absolute calibration, while standard IAA metrics demand exact numerical consensus and are highly susceptible to statistical deflation under skewed distributions.

To address these gaps, we introduce Tolerance-Weighted Agreement (TWA, ω), an absolute alignment metric natively designed for the LLM-as-a-Judge paradigm. By using graded tolerance weights to absorb minor subjective variance, TWA isolates true semantic consensus between artificial judges and human evaluators without suffering from statistical deflation. We validate the metric across an In-Context Learning (ICL) classification explainability dataset and the evaluation sets of established benchmarks (BiGGen-Bench, SummEval). The results demonstrate TWA's robustness in quantifying Human-LLM alignment across diverse tasks. Furthermore, it provides a highly interpretable alternative to complex metrics like Gwet's AC2 and naturally scales to measure Inter-Judge Agreement directly between competing LLMs.

1 INTRODUCTION

The paradigm of utilizing Large Language Models (LLMs) as judges has emerged as a rapidly growing practice, offering unprecedented speed and scalability for evaluating complex tasks. However, the field currently lacks precise statistical methodologies to accurately quantify how well these artificial judges align with human consensus, especially on granular scales (e.g., 1-5 or 1-10 point Likert rubrics).

To resolve this gap, we propose Tolerance-Weighted Agreement (TWA, denoted as ω). Grounded in a Discrete Normal Distribution, TWA safely absorbs natural human subjective variance (e.g., the difference between a 4 and a 5) without relying on flawed chance-agreement penalties, providing a reliable and scale-invariant standard for measuring true evaluator alignment.

Our main contributions are:

- **Identification of the Kappa Paradox in LLM Evaluation:** We empirically demonstrate that current evaluation paradigms relying on rank correlations (which ignore absolute calibration) or traditional chance-adjusted metrics (which suffer from large statistical deflation on skewed datasets, known as the *Kappa Paradox* (Feinstein & Cicchetti, 1990)) are fundamentally flawed for granular evaluation.
- **Tolerance-Weighted Agreement (TWA):** We introduce a novel, absolute alignment metric grounded in Gaussian tolerance that robustly tracks consensus without statistical deflation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Comprehensive Empirical Validation:** We demonstrate TWA's robustness across different LLM-as-a-Judge evaluation datasets, establishing its capacity to evaluate both Human-LLM consensus and direct Inter-Judge Alignment.
- **Thresholds:** We establish operational alignment thresholds ($\omega \geq 0.75$) mathematically proven to mandate a tighter-than-human evaluative variance ($\sigma \leq 0.85$) to certify artificial judges.
- **Open-Source Python Library:** To bridge the gap between theoretical rigor and practical application, a lightweight open-source Python library is provided to allow researchers to readily integrate TWA into existing evaluation workflows.

2 Background and Problem Formulation

2.1 Linear and Rank Correlation Metrics

In the current LLM-as-a-Judge literature, correlation coefficients—such as Pearson's r (Pearson, 1895), or rank metrics like Spearman's ρ and Kendall's τ (Kendall, 1938)—are standard evaluation metrics. For instance, FLASK (Ye et al., 2024) relies on Pearson and Kendall's τ to validate alignment, while G-Eval (Liu et al., 2023) uses Spearman to show that GPT-4 aligns with human judgment. However, these metrics are unsuited for granular evaluation scales (e.g., 1-5 Likert rubrics) because they do not account for absolute calibration. A correlation coefficient can yield a perfect score (1.0) even if a human assigns scores of (3, 4, 5) while a systematically strict LLM assigns (1, 2, 3) to the same outputs. This focus on relative order can obscure large discrepancies in absolute scoring.

2.2 Inter-Annotator Agreement (IAA) Metrics and the Kappa Paradox

To measure absolute alignment rather than relative order, the traditional standard dictates the use of Inter-Annotator Agreement metrics such as Cohen's Kappa (κ), Weighted Kappa (Cohen, 1968), and Krippendorff's Alpha (α). Cohen's Kappa measures exact numerical agreement while penalizing for chance agreement (p_e). Weighted Kappa and Krippendorff's Alpha similarly use distance metrics to partially forgive minor deviations (e.g., a 4 vs a 5); the latter is used by SummEval (Fabbri et al., 2021) to measure human consensus.

However, when applied to modern LLM evaluation, these metrics exhibit a well-documented statistical limitation known as the Kappa Paradox (Feinstein & Cicchetti, 1990) or Kappa Deflation. Because modern LLMs consistently generate high-quality text, evaluation datasets are typically skewed towards the upper bounds of the scale (e.g., predominantly 4s and 5s). Since Kappa and Alpha are chance-adjusted metrics, this high prevalence of top scores inflates the expected mathematical probability of chance agreement (p_e). As a result, these metrics report disproportionately low alignment coefficients even when the absolute agreement (raw agreement) between evaluators exceeds 90%. This deflation is prevalent in modern NLP benchmarks; for instance, the SummEval benchmark (Fabbri et al., 2021) reported a large drop in Inter-Annotator Agreement (IAA), yielding a Krippendorff's Alpha of just 0.41 among expert human annotators. The unreliability of these metrics under skewed distributions highlights that traditional IAA metrics are unsuited for this task.

2.3 Gwet's AC2: A Solid Alternative from Biomedicine

To address the limitations of traditional Kappa metrics under skewed distributions, Gwet's AC2 (Gwet, 2014) has emerged as the most solid and robust chance-adjusted alternative. AC2 recalibrates the chance-agreement probability to prevent the deflation paradox, making it highly reliable for ordinal data with skewed marginals. However, this metric is not frequently seen in Artificial Intelligence (AI) and Natural Language Processing (NLP) publications; rather, it firmly belongs to the domain of biomedicine and clinical research.

While our experimentation will show a strong correlation between AC2 and our proposed metric (see Section 4), a fundamental limitation of AC2 for the LLM-as-a-Judge paradigm is its strict dependence on global marginal frequencies. AC2 adjusts for chance agreement using a prevalence-anchored weighted probability P_e^w, which inherently requires calculating the marginal probability π_k of every score category k across the entire dataset (see Appendix J for the exact mathematical

formulation). Because of this global dependency, the metric cannot be processed iteratively. If a new evaluation arrives, the dataset denominator and the global probabilities change, retroactively altering the chance-agreement penalty and yielding an abstract reliability coefficient that obscures operational meaning.

In contrast, TWA addresses this gap by operating entirely independent of dataset-level distributions. By explicitly modeling natural subjectivity through Gaussian distance rather than adjusting for random guessing, the formulation of TWA depends exclusively on the distance at the current instance i:

$$\omega_i = W(|y^{(i)}_{judge} - y^{(i)}_{human}|)$$

This absolute locality mathematically anchors the metric as an interpretable "relaxed accuracy." Because it operates independently of global marginals, TWA can be calculated incrementally, which is a crucial architectural advantage for real-time evaluation pipelines.

3 Tolerance-Weighted Agreement (ω)

3.1 Metric Formulation and Rationale

Let $y^{(i)}_{human}, y^{(i)}_{judge} \in \{1, 2, 3, 4, 5\}$ represent the scores assigned by the human and the algorithmic judge for the i-th evaluated instance, respectively. Defining the absolute distance between the judge and human scores as $d_i = |y^{(i)}_{judge} - y^{(i)}_{human}|$, the tolerance weight function $W(d)$ is derived by discretizing a standard Normal distribution $\mathcal{N}(0, 1)$[1]. For each integer distance d, the weight is the probability mass of the bin $[d - 0.5,\ d + 0.5]$ normalized by the mass of the central bin $[-0.5,\ 0.5]$:

$$W(d) = \frac{\Phi(d + 0.5) - \Phi(d - 0.5)}{\Phi(0.5) - \Phi(-0.5)}$$

where $\Phi(\cdot)$ denotes the standard Normal CDF. This yields a parameter-free, principled weight for each distance level (see Appendix A, Figure 5):

Table 1: Tolerance weights $W(d)$ derived from Gaussian Discretization.

Distance d	**Weight** $W(d)$	**Interpretation**
0	1.000	Perfect agreement
1	0.631	Acceptable subjective variance
2	0.158	Marginal divergence
≥ 3	≈ 0	large misalignment

The overall TWA score is:

$$\omega = TWA = \frac{1}{N} \sum_{i=1}^{N} W(d_i)$$

The weight at $d = 1$ (≈ 0.63) is not an arbitrary choice but the exact proportion of probability mass assigned by the standard Normal to a one-unit deviation, normalized against perfect agreement.

To interpret these results operationally, we define categorical alignment labels with thresholds recalibrated to account for the systematically lower—yet mathematically stricter—weights produced

[1]We fix $\sigma = 1.0$ as a universal reference convention, deliberately independent of any dataset's empirical variance—analogous to how conventional effect-size thresholds (e.g., Cohen's d) remain fixed across domains to preserve cross-study comparability. The empirical σ we observe varying by dimension (Appendix G, $\sigma \in [0.6, 2.04]$) is not a violation of this assumption but a diagnostic signal: it reveals which evaluation dimensions carry genuinely higher subjective noise, information that a dataset-adaptive metric would normalize away.

by the Gaussian discretization:

$$Alignment(\omega) = \begin{cases} \text{Aligned} & \text{if } \omega \geq 0.75 \\ \text{Partially Aligned} & \text{if } 0.65 \leq \omega < 0.75 \\ \text{Misaligned} & \text{if } \omega < 0.65 \end{cases} \quad (1)$$

By definition, perfect agreement yields a TWA score of 1.0. Crucially, because $W(1) \approx 0.631$, an artificial judge that never achieves exact agreement but is consistently off by 1 point will score 0.631, thus correctly falling into the **Misaligned** category. This ensures that an LLM cannot be considered aligned merely by being "close"; it must achieve a baseline frequency of exact matches to reach the Partial Alignment threshold (0.65). We establish $\omega \geq 0.75$ as a reference threshold for strong alignment; while not an absolute truth, it provides a practical basis for concrete decisions. An **Aligned** score ($\omega \geq 0.75$) indicates sufficient consensus for formal benchmarking, while a **Partially Aligned** score ($\omega \geq 0.65$) indicates that large disagreements ($|d| \geq 2$) remain rare even though exact point-matching fluctuates—adequate for refining rubrics and judge prompts, though not for final ground-truth scoring. We propose these thresholds as a standardized baseline for future LLM-as-a-Judge research.

Confidence Intervals and Statistical Stability. A mathematical property of TWA over traditional chance-adjusted metrics (such as Cohen's κ, Krippendorff's α, or Gwet's AC2) is its statistical simplicity. Because traditional metrics involve non-linear chance adjustments, calculating their variance requires complex covariance matrices and often lacks exact solutions for finite samples. In contrast, TWA is a simple sample mean of independent instance-level weights. By the Central Limit Theorem, its Standard Error (SE) is directly estimated as $SE = \frac{s_W}{\sqrt{N}}$ (where s_W is the sample standard deviation of the assigned weights), and the 95% Confidence Interval can be calculated as $CI = \omega \pm 1.96 \cdot SE$. For the benchmark datasets evaluated in this work, empirical standard errors consistently remain low (e.g., $SE \leq 0.02$), ensuring that the reported TWA scores are stable and statistically significant without requiring complex bootstrapping procedures.

4 Experimental Evaluation

4.1 Experimental Setup and Evaluation Datasets

We validate the Tolerance-Weighted Agreement (TWA) framework starting with an exhaustive analysis of an In-Context Learning (ICL) explainability dataset introduced by Quiles-Ramírez et al. (Quiles-Ramírez et al., 2026). In this setting, LLM explainability is evaluated across 9 dimensions on a 1-5 Likert scale, utilizing both LLM-as-a-Judge and human expert annotations. To minimize methodological misalignment, this dataset ensures that both human and artificial judges operated under the same explicitly anchored rubrics. To demonstrate the metric's broader applicability, we subsequently extend our evaluation to the datasets from two established benchmarks: BiGGen-Bench (Kim et al., 2025) (Prometheus-2-8x7B (Kim et al., 2024) judge) and SummEval (Fabbri et al., 2021) (G-Eval/GPT-4 (Liu et al., 2023) judge), specifically utilizing their paired human and LLM judge results. We intentionally exclude MT-Bench (Zheng et al., 2023) from this analysis; because its pairwise win/tie/loss format effectively operates on a narrow 3-point scale ($N = 3$), it falls outside the scope of our study. We also excluded FLASK (Ye et al., 2024): the only available paired human-LLM dataset was a third-party repackaging we could not verify against the official release (Appendix F).

We evaluate TWA (ω) against a comprehensive suite of established evaluation metrics. For absolute measurement, we report Exact Match as a strict absolute baseline and Mean Absolute Error (MAE) to quantify average error magnitude. To assess chance-corrected agreement and demonstrate Kappa Deflation, we utilize Cohen's Kappa (κ) (Cohen, 1960), Weighted Kappa (Cohen, 1968), interval-based Krippendorff's Alpha (α) (Hayes & Krippendorff, 2007), and Gwet's AC2 (Gwet, 2014). For correlation, we report Pearson's r (Pearson, 1895) for linear relationships, alongside Spearman's ρ (Spearman, 1904) and Kendall's τ (Kendall, 1938) for rank-based evaluation.

4.2 USING TWA FOR ALIGNMENT EVALUATION: LLM JUDGE VS. HUMAN

We evaluate TWA using a granular explainability research dataset (Quiles-Ramírez et al., 2026), where a state-of-the-art LLM judge and independent human experts assessed generated text explanations across nine distinct evaluation dimensions.

Table 2 details the error distributions and TWA scores for this dataset, revealing calibration differences depending on task complexity. For straightforward dimensions like Conciseness and Comprehensibility, TWA approaches or exceeds 0.90, indicating high agreement. Conversely, for abstract tasks like Discriminativeness, TWA drops below 0.70, highlighting where the LLM struggles to capture human subjective reasoning. Across all dimensions, most deviations are isolated to ± 1. This rapid decay in disagreement tightly follows a Discrete Normal distribution (see Appendix F, Figure 10 for a detailed KL Divergence analysis). Such a pattern implies that large evaluative misalignments ($|d| \geq 2$) are uncommon, and that the LLM's scoring variance closely mirrors natural human subjectivity.

Table 2: Distribution of Error Magnitudes. Exact match and ± 1 tolerances comprise the TWA score. Aligned scores (TWA ≥ 0.75) are in **bold**.

Evaluation Dimension	Exact Match (0)	+1 Diff	-1 Diff	MAE	TWA	Alignment
Conciseness	85.9%	7.0%	1.0%	0.21	**0.918**	Aligned
Comprehensibility	81.5%	3.9%	7.6%	0.29	**0.894**	Aligned
Hallucination Free	75.8%	3.1%	7.3%	0.49	**0.834**	Aligned
Logical Coherence	63.5%	13.8%	8.3%	0.61	**0.787**	Aligned
Instruction Following	64.1%	6.2%	12.2%	0.58	**0.779**	Aligned
Concept Counting	60.4%	13.0%	2.6%	0.87	0.720	Partially Aligned
Textual Groundedness	41.4%	9.1%	35.2%	0.77	0.711	Partially Aligned
Specificity	46.9%	13.3%	20.8%	0.77	0.708	Partially Aligned
Discriminativeness	45.3%	9.6%	15.9%	0.98	0.644	Misaligned

For subjective dimensions like *Comprehensibility*, the ± 1 tolerance margins contribute substantially to the alignment score, bridging rigid exactness and acceptable human variance. Importantly, TWA does not artificially inflate scores for objective tasks: in *Concept Counting*, where a 1-point deviation represents a factual error, TWA correctly assigns a lower alignment score (0.720).

Tables 3 and 4 compare TWA against traditional statistical metrics, empirically exposing their inherent limitations. In both tables, Exact Match and Mean Absolute Error (MAE) are included as grounded reference points for the actual magnitude of error. As shown in Table 3, rank correlations (Spearman's ρ, Kendall's τ (Kendall, 1938)) and linear correlations (Pearson's r (Pearson, 1895)) fail to capture absolute calibration. For instance, Spearman's ρ scores the severely misaligned *Discriminativeness* dimension ($\rho = 0.55$, 45.3% Exact Match, MAE 0.98) higher than the firmly aligned *Logical Coherence* ($\rho = 0.54$, 63.5% Exact Match, MAE 0.61). TWA, by contrast, accurately reflects the structural deterioration indicated by Exact Match and MAE.

Table 3: Correlation Metrics vs. Tolerance-Weighted Agreement (TWA) on the ICL dataset. Rank correlations (Spearman, Kendall) and linear correlation (Pearson) often obscure absolute rating differences, whereas TWA natively penalizes absolute calibration errors. Sig. indicates statistical significance ($p_{adj} < 0.05$) against a random marginal baseline.

Evaluation Dimension	Exact Match	MAE	Pearson r	Spearman ρ	Kendall τ	TWA (ω) (Alignment)	Sig.
Conciseness	85.9%	0.21	0.70	0.53	0.51	**0.92** (Aligned)	✓
Comprehensibility	81.5%	0.29	0.68	0.67	0.65	**0.89** (Aligned)	✓
Hallucination Free	75.8%	0.49	0.35	0.34	0.32	**0.83** (Aligned)	✓
Logical Coherence	63.5%	0.61	0.53	0.54	0.49	**0.79** (Aligned)	✓
Instruction Following	64.1%	0.58	0.62	0.64	0.57	**0.78** (Aligned)	✓
Concept Counting	60.4%	0.87	0.22	0.16	0.15	0.72 (Partially Aligned)	✓
Textual Groundedness	41.4%	0.77	0.71	0.64	0.57	0.71 (Partially Aligned)	✓
Specificity	46.9%	0.77	0.62	0.53	0.47	0.71 (Partially Aligned)	✓
Discriminativeness	45.3%	0.98	0.56	0.55	0.46	0.64 (Misaligned)	✓

Furthermore, Table 4 illustrates the mathematical collapse of chance-adjusted agreement metrics. The Kappa Deflation phenomenon is most evident in the *Conciseness* dimension: despite an 85.9% Exact Match and a 0.918 TWA score, Cohen's κ collapses to 0.28 (merely 'Fair'). While more sophisticated metrics like Weighted κ (Cohen, 1968) (0.61) and Krippendorff's interval α (0.61) attempt to adjust for chance or ordinality, they still yield highly fluctuating or conflicting reliability scores across di-

mensions. Recent literature evaluating LLMs in specialized domains has recognized the susceptibility

Table 4: Chance-Adjusted Agreement Metrics vs. TWA on the ICL dataset. Due to *Kappa Deflation* under skewed marginal distributions, traditional metrics (Cohen's κ, Weighted κ, Interval α, Gwet's AC2) often collapse or yield highly contradictory reliability scores. TWA provides a stable, absolute measure of consensus. Sig. indicates statistical significance ($p_{adj} < 0.05$) against a random marginal baseline.

Evaluation Dimension	Exact Match	MAE	Cohen's κ	Weighted κ	Interval α	Gwet's AC2	TWA (ω)	Sig.
Conciseness	85.9%	0.21	0.28	0.61	0.61	0.97	**0.92**	✓
Comprehensibility	81.5%	0.29	0.49	0.59	0.59	0.94	**0.89**	✓
Hallucination Free	75.8%	0.49	0.24	0.34	0.34	0.88	**0.83**	✓
Logical Coherence	63.5%	0.61	0.33	0.53	0.53	0.83	**0.79**	✓
Instruction Following	64.1%	0.58	0.38	0.61	0.61	0.85	**0.78**	✓
Concept Counting	60.4%	0.87	0.10	0.22	0.22	0.75	0.72	✓
Textual Groundedness	41.4%	0.77	0.27	0.66	0.65	0.73	0.71	✓
Specificity	46.9%	0.77	0.32	0.60	0.60	0.72	0.71	✓
Discriminativeness	45.3%	0.98	0.27	0.55	0.54	0.56	0.64	✓

of traditional metrics to these skewed distributions, actively advocating for paradox-resistant alternatives like Gwet's AC2 (Gwet, 2014; Pradhan et al., 2025). Our empirical results corroborate this: Gwet's AC2 successfully stabilizes against deflation (0.97 for *Conciseness*). However, TWA (0.918) achieves mathematically equivalent robustness while maintaining direct, actionable interpretability over ordinal distances, unlike the abstract probability corrections underlying AC2.

4.3 Using TWA for Inter-Annotator Agreement: Human vs. Human

To further validate that the LLM's subjective calibration mirrors natural human variance, we can effectively extend the TWA formula to measure absolute Inter-Annotator Agreement between two independent humans. By replacing the algorithmic judge's prediction with a second human's prediction, the distance function becomes:

$$d_i = |y^{(i)}_{human1} - y^{(i)}_{human2}| \tag{2}$$

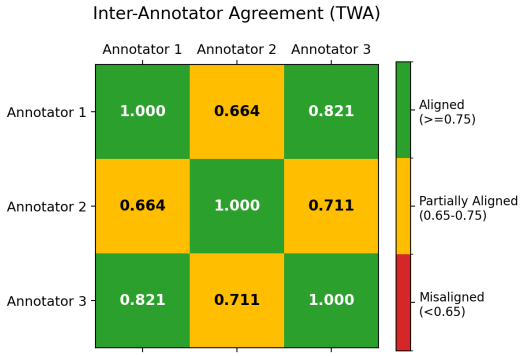


Figure 1: Inter-Annotator Alignment (TWA) Matrix. The 3x3 heatmap displays the pairwise TWA consensus between the three independent human annotators across all dimensions.

Figure 1 demonstrates that the three independent human annotators maintain a strong TWA consensus amongst themselves, which mirrors the alignment scores achieved by the LLM-as-a-judge. This confirms that the variance absorbed by the $d = 1$ tolerance window is a natural property of subjective evaluation, and that the LLM operates indistinguishably from a human peer in this regard. For a highly detailed breakdown mapping individual annotators to specific evaluation dimensions, see Appendix F.

4.4 Evaluating TWA on SOTA Benchmark Datasets

We now extend our experimentation to the evaluation datasets from widely adopted State-of-the-Art (SOTA) benchmarks (BiGGen-Bench and SummEval), utilizing their paired human and LLM judge

results. When applying Tolerance-Weighted Agreement, we observe a substantial recalibration in the alignment metrics (see Tables 5 and 6). For instance, on the BiGGen-Bench dataset (Prometheus-2-8x7B judge), Exact Match yields baseline agreements of 26.3%–41.1%, while TWA reveals empirical consensuses of 0.569–0.705. In SummEval, TWA ranges from 0.196 (for the severely misaligned Fluency dimension) up to 0.694 (for Relevance). Furthermore, a permutation test ($K = 10,000$) reveals that G-Eval's Fluency scores exhibit an extreme negative z-score ($z = -7.32$, $p_{adj} = 1.0000$ in the upper-tail test), demonstrating significant systematic discordance (severe under-scoring bias) rather than random agreement. This proves TWA isolates pure evaluative noise without requiring complex chance-adjusted formulas (see Appendix I for full significance testing).

Table 5: Correlation Metrics vs. Tolerance-Weighted Agreement (TWA) on SOTA Benchmarks (BiGGen-Bench and SummEval). Missing correlations indicate datasets where variance was zero (all identical human scores) preventing valid rank calculations. Sig. indicates statistical significance ($p_{adj} < 0.05$) against a random marginal baseline.

Dataset	Evaluation Dimension	Exact Match	MAE	Pearson r	Spearman ρ	Kendall τ	TWA (ω) (Alignment)	Sig.
BiGGen-Bench (Kim et al., 2025)	Grounding	39.8%	0.80	0.59	0.55	0.50	0.70 (Partially Aligned)	✓
BiGGen-Bench (Kim et al., 2025)	Safety	41.1%	0.81	0.57	0.49	0.43	0.70 (Partially Aligned)	✓
BiGGen-Bench (Kim et al., 2025)	Planning	38.2%	0.79	0.59	0.57	0.50	0.70 (Partially Aligned)	✓
BiGGen-Bench (Kim et al., 2025)	Theory Of Mind	35.2%	0.86	0.39	0.37	0.33	0.67 (Partially Aligned)	✓
BiGGen-Bench (Kim et al., 2025)	Reasoning	35.0%	0.90	0.56	0.54	0.47	0.66 (Partially Aligned)	✓
BiGGen-Bench (Kim et al., 2025)	Instruction Following	28.5%	0.93	0.48	0.41	0.36	0.65 (Misaligned)	✓
BiGGen-Bench (Kim et al., 2025)	Tool Usage	31.2%	1.02	0.46	0.44	0.38	0.62 (Misaligned)	✓
BiGGen-Bench (Kim et al., 2025)	Refinement	26.3%	1.14	0.45	0.48	0.41	0.57 (Misaligned)	✓
SummEval (Fabbri et al., 2021)	Relevance	36.7%	0.79	0.53	0.51	0.46	0.69 (Partially Aligned)	✓
SummEval (Fabbri et al., 2021)	Coherence	34.8%	0.86	0.52	0.53	0.47	0.67 (Partially Aligned)	✓
SummEval (Fabbri et al., 2021)	Consistency	27.4%	1.07	0.54	0.48	0.44	0.60 (Misaligned)	✓
SummEval (Fabbri et al., 2021)	Fluency	2.5%	2.05	0.54	0.47	0.45	0.20 (Misaligned)	✗

Table 6: Chance-Adjusted Agreement Metrics vs. TWA on SOTA Benchmarks. The degradation of Cohen's κ is evident across almost all dimensions, whereas TWA provides a stable measure of consensus. Sig. indicates statistical significance ($p_{adj} < 0.05$) against a random marginal baseline.

Dataset	Evaluation Dimension	Exact Match	MAE	Cohen's κ	Weighted κ	Interval α	Gwet's AC2	TWA (ω)	Sig.
BiGGen-Bench (Kim et al., 2025)	Grounding	39.8%	0.80	0.22	0.54	0.54	0.73	0.70	✓
BiGGen-Bench (Kim et al., 2025)	Safety	41.1%	0.81	0.21	0.55	0.54	0.72	0.70	✓
BiGGen-Bench (Kim et al., 2025)	Planning	38.2%	0.79	0.15	0.50	0.49	0.79	0.70	✓
BiGGen-Bench (Kim et al., 2025)	Theory Of Mind	35.2%	0.86	0.11	0.33	0.33	0.76	0.67	✓
BiGGen-Bench (Kim et al., 2025)	Reasoning	35.0%	0.90	0.19	0.51	0.51	0.68	0.66	✓
BiGGen-Bench (Kim et al., 2025)	Instruction Following	28.5%	0.93	0.06	0.41	0.41	0.74	0.65	✓
BiGGen-Bench (Kim et al., 2025)	Tool Usage	31.2%	1.02	0.14	0.41	0.41	0.60	0.62	✓
BiGGen-Bench (Kim et al., 2025)	Refinement	26.3%	1.14	0.10	0.34	0.29	0.53	0.57	✓
SummEval (Fabbri et al., 2021)	Relevance	36.7%	0.79	0.11	0.39	0.30	0.80	0.69	✓
SummEval (Fabbri et al., 2021)	Coherence	34.8%	0.86	0.13	0.38	0.30	0.74	0.67	✓
SummEval (Fabbri et al., 2021)	Consistency	27.4%	1.07	0.07	0.35	0.21	0.70	0.60	✓
SummEval (Fabbri et al., 2021)	Fluency	2.5%	2.05	-0.02	0.09	-0.56	0.22	0.20	✗

Somewhat surprisingly, no external-benchmark dimension reaches the Aligned category. The original authors of BiGGen-Bench (Kim et al., 2025) and SummEval (Fabbri et al., 2021) both documented notable difficulties in their human evaluation protocols, and we read the observed TWA values largely as a reflection of those documented difficulties rather than as a verdict on the judges themselves. TWA accounts for these difficulties by mathematically separating minor subjective noise ($d = 1$) from large evaluative failures ($d \geq 2$). Furthermore, our analysis (Appendix B, Figure 11) shows these error variances closely fit a Discrete Normal distribution, indicating that LLM evaluations primarily exhibit standard Gaussian variance rather than random noise.

4.5 LLM-to-LLM Alignment (Inter-Judge Agreement)

A significant advantage of TWA is its ability to measure alignment not only between humans and LLMs, but directly between competing LLM judges without requiring human ground truth. Using the BiGGen-Bench dataset, which includes evaluations from Prometheus-8x7B, GPT-4, and Claude, we compute the Inter-Judge Agreement across all dimensions.

4.6 Threshold Justification and Deterministic Bounds

We establish the threshold for strong alignment at $\omega \geq 0.75$ because it provides a mathematical guarantee of high reliability while accommodating natural LLM variance.

Furthermore, to address chance agreement without relying on flawed Kappa penalties, the theoretical lower bound of TWA for a uniformly random judge on a 1-5 scale is strictly bounded at $\omega \approx 0.44$. While this establishes a strict mathematical floor, practical empirical distributions are naturally dictated by the dataset's rubric and marginal skew. TWA correctly reports high alignment when

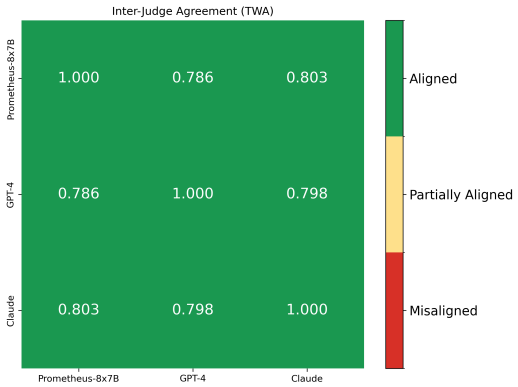


Figure 2: Inter-Judge Agreement (TWA) Matrix on BiGGen-Bench. The heatmap displays the pairwise consensus between Prometheus-8x7B, GPT-4, and Claude.

judges consistently agree on highly prevalent scores. To confirm that this consensus stems from true evaluation rather than defaulting to the most frequent score, TWA is deployed alongside the permutation test (Section 4.4). This ensures the evaluator is actively responding to the semantic content ($p_{adj} < 0.05$).

A complete analysis of TWA's operational meaning under these deterministic edge-case error profiles is provided in Appendix K.

4.7 Numerical Simulation: TWA vs. Gwet's AC2

To rigorously evaluate TWA's theoretical behavior against Gwet's AC2 (a leading paradox-resistant ordinal metric), we examine two critical statistical properties: its simulated response to marginal skew, and its analytical penalty profile for extreme errors.

Sensitivity to Systematic Bias (Shifted Error): A critical flaw of traditional metrics is that their global formulations obscure specific directions of annotator bias. Structurally, TWA calculates consensus strictly point-by-point ($d_i = |y_1^{(i)} - y_2^{(i)}|$), responding deterministically to the absolute magnitude of the error. To test how TWA and AC2 respond to a systematically biased judge, we simulated an evaluation pipeline of 10,000 instances. We fixed the injected noise standard deviation at $\sigma = 1.0$, but shifted the mean of the error (μ_{error}) from -3.0 (overly harsh) to $+3.0$ (overly lenient) (see Appendix H). As shown in Figure 3 (left), both TWA and AC2 successfully capture this systematic degradation symmetrically, forming an inverted-V shape that plummets as the bias increases. This proves that TWA robustly drops when a model systematically hallucinates in one direction.

Sensitivity to Extreme Errors: A robust LLM evaluation metric must severely penalize large misalignments ($|d| \geq 2$) while forgiving minor subjective noise ($|d| \leq 1$). Rather than simulating variance, this property is best isolated analytically through the metrics' distance penalty functions. As shown in Figure 3 (right), TWA accomplishes this natively because it relies on a Gaussian decay ($\omega_{i,j} \propto \exp(-d^2/2)$) rather than AC2's forgiving quadratic penalty ($1 - (d/4)^2$). TWA is significantly more sensitive to extreme errors: a distance of $d = 2$ on a 5-point scale retains 75% partial credit under AC2, but only 13.5% under TWA. When judges produce large deviations, TWA correctly drops to reflect the lack of safety, whereas AC2's inflated partial credit keeps the metric overly optimistic.

5 Limitations and Future Work

While Tolerance-Weighted Agreement (ω) provides a robust framework for LLM-as-a-Judge evaluation, our current empirical validation is primarily constrained to 5-point Likert scales. These odd-numbered scales possess a natural neutral center, supporting symmetric Gaussian discretization. The mathematical behavior of TWA on even-numbered, forced-choice rubrics (e.g., 4-point or 6-point)

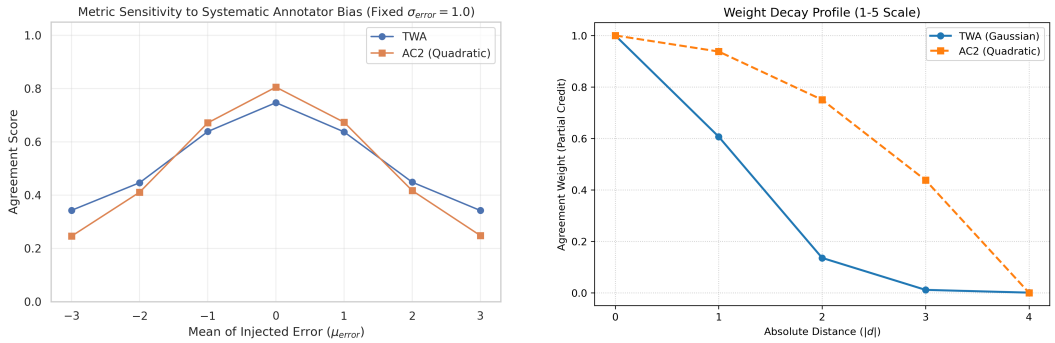


Figure 3: Theoretical comparison between TWA and AC2. **Left:** Numerical Simulation ($N = 10,000$) measuring sensitivity to systematic annotator bias. The mean of the injected Gaussian error is shifted from -3.0 to +3.0 while fixing $\sigma = 1.0$. Both metrics cleanly capture this degradation with a symmetric inverted-V curve. **Right:** Analytical weight decay profiles. TWA's Gaussian formulation severely penalizes extreme errors ($|d| \geq 2$) compared to AC2's forgiving quadratic weights, ensuring that large misalignments correctly plunge the metric score.

remains an unexplored area requiring targeted validation. Furthermore, while we propose a theoretical adaptation for 10-point scales, extensive experimentation on continuous or highly granular scales (e.g., 1-100) is necessary to verify if a standard normal distribution remains the optimal prior for subjective variance. Conversely, TWA degrades on highly compressed rubrics (e.g., 3-point scales), where a $d \leq 1$ tolerance absorbs too much of the total variance to remain discriminative.

TWA also assumes zero-centered disagreement; as shown in Appendix F, a judge can exhibit a systematic scoring bias rather than symmetric noise. Future work should explore bias-aware extensions—such as localized residual corrections—while preserving instance-level efficiency.

6 Conclusion

As Large Language Models increasingly serve as autonomous judges, evaluating them with strict Exact Match, rank correlations (e.g., Spearman's ρ), or chance-adjusted metrics like Cohen's Kappa systematically mischaracterizes their true capabilities. To address this, we introduced Tolerance-Weighted Agreement (TWA), a principled metric that resolves Kappa Deflation by explicitly modeling human subjective variance as a discrete Gaussian distribution.

Through extensive empirical validation across three granular datasets—a custom explainability study alongside paired results from BiGGen-Bench and SummEval—we demonstrated that TWA isolates natural subjective noise from large evaluative misalignment. Crucially, rigorous non-parametric permutation tests confirmed that TWA acts as a robust statistical filter, distinguishing true evaluative consensus from models randomly guessing according to their marginal distributions. Furthermore, we demonstrated that TWA provides statistically stable confidence intervals and can directly measure Inter-Judge Agreement between competing LLMs in the absence of human annotations. By establishing statistically justified operational thresholds, TWA equips researchers to confidently deploy and calibrate artificial judges without falling victim to chance agreement or penalizing subjective human variance.

AI DISCLOSURE

In this work, we used generative AI tools for editing the research paper to improve readability, translating components of the text, writing and formatting software code for data extraction, generating and formatting scientific figures, and resolving formatting errors in the LaTeX manuscript. We have not used generative AI tools for generating synthetic datasets, formulating mathematical claims, proving mathematical claims, proposing hypotheses, or interpreting results, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work, including all software code which was verified and tested for correctness by the authors, and manual verification of all LaTeX changes. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on the methodological evaluation of Large Language Models (LLMs) as automated judges. We utilize publicly available datasets and previously published human annotations (e.g., BiGGen-Bench, SummEval) to conduct our analysis. No new human subjects were recruited, and no sensitive or personally identifiable information (PII) was collected or processed during this study. By providing a more statistically rigorous metric (TWA) to evaluate LLM alignment with human consensus, this research aims to mitigate the risks of deploying biased or miscalibrated automated evaluators in real-world scenarios. We acknowledge that any automated evaluation metric, including TWA, should be used as a complement to—rather than a complete replacement for—diverse human oversight in high-stakes domains.

REFERENCES

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

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

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

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

Kilem L. Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, 4th edition, 2014.

Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, et al. 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 (EMNLP)*, 2024.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, et al. The BiGGen Bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

Anu Pradhan, Alexandra Ortan, Apurv Verma, and Madhavan Seshadri. Llm-as-a-judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation. In *Proceedings of the 2nd Workshop on Evaluating and Applying Recommender Systems with Large Language Models (EARL) at RecSys*, 2025.

Carmen Quiles-Ramírez, Leticia L. Rodríguez, Nicolás Martorell, and Natalia Díaz-Rodríguez. Explaining is harder than predicting alone: Evaluating concept-based explanations of MLLMs as ICL visual classifiers. In *ICML CompLearn Workshop*, 2026.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, et al. Flask: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) 36, Datasets and Benchmarks Track*, 2023.

A HEURISTIC PREDECESSOR: TWA STEP-FUNCTION

The Tolerance-Weighted Agreement (TWA) framework introduced in Section 3 extends this formulation, replacing the hardcoded step-function (Quiles-Ramírez et al. (2026)) with weights derived analytically from a discrete Gaussian distribution. Given the absolute distance $d_i = |y^{(i)}_{judge} - y^{(i)}_{human}|$, the item-level scoring function was defined as:

$$S(d_i) = \begin{cases} 1 & \text{if } d_i = 0 \\ 0.75 & \text{if } d_i = 1 \\ 0 & \text{if } d_i > 1 \end{cases}$$

Their overall relaxed accuracy was then the mean across all N instances: $Score = \frac{1}{N}\sum_{i=1}^{N} S(d_i)$.

The rationale for $S(1) = 0.75$ was probabilistic: assuming that inter-annotator variance follows a discrete Gaussian $d \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 1/1.15$, approximately 75% of the probability mass concentrates within a 1-point margin (Figure 4). The corresponding alignment thresholds were: **Aligned** ≥ 0.75, **Partially Aligned** ≥ 0.65, and **Misaligned** < 0.65.

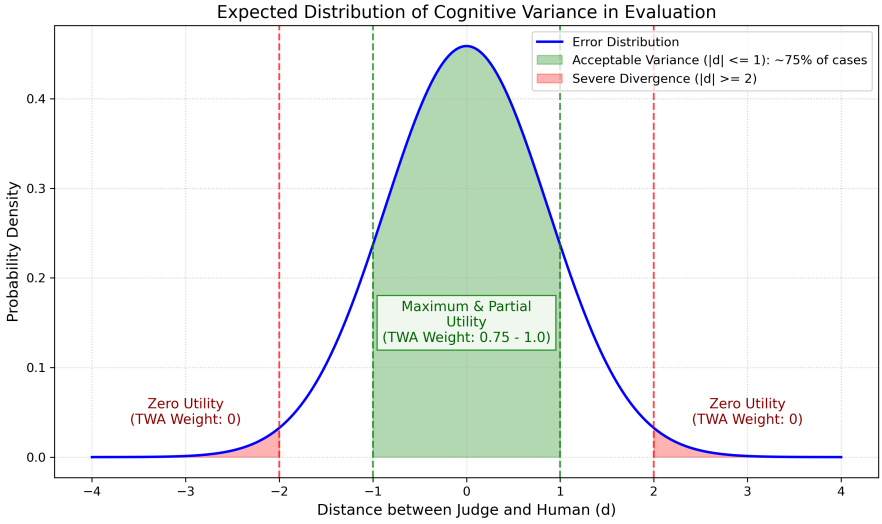


Figure 4: Expected theoretical distribution of subjective variance justifying the original $S(1) = 0.75$ step-function weight. The ± 1 threshold captures $\approx 75\%$ of natural subjective calibration noise.

While this formulation is operationally valid and interpretable, it has two limitations: (1) the 0.75 weight is an approximation derived from a specific σ choice, and (2) the hard zero at $d \geq 2$ discards all signal from marginal divergences. The discrete Gaussian formulation adopted in this paper (Section 3) resolves both issues. It naturally accommodates any n-point Likert scale without re-tuning parameters, and the continuous decay ensures that marginal divergences ($d = 2$) receive small but non-zero credit, while large errors ($d \geq 3$) are effectively penalized with near-zero weight.

Discrete Gaussian Derivation As shown in Figure 5, the weight $W(d) = \frac{\Phi(d+0.5)-\Phi(d-0.5)}{\Phi(0.5)-\Phi(-0.5)}$ extracts the exact probability mass of each discrete bin relative to the central bin, producing $W_0 = 1.00$, $W_1 \approx 0.63$, $W_2 \approx 0.16$, with a near-zero tail for $d \geq 3$. The original alignment thresholds (Aligned ≥ 0.75, Partially Aligned ≥ 0.65) are preserved unchanged and are now stricter in practice, since the Gaussian weights are lower.

B ADAPTIVE SCALING FOR N-POINT LIKERT SCALES

To adapt the TWA metric to larger granular scales (e.g., 1-10) while preserving the same statistical tolerance limits, the underlying Gaussian distribution must scale proportionally.

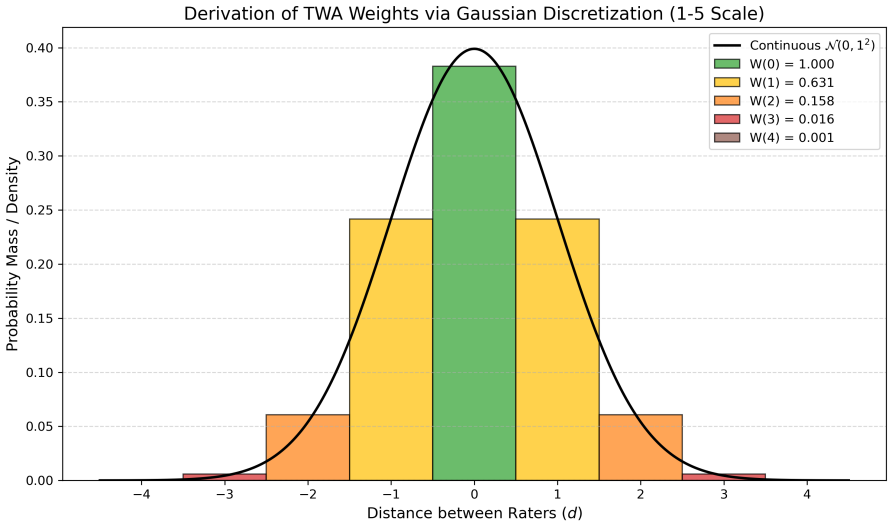


Figure 5: Derivation of TWA weights via Gaussian Discretization. Dynamic tolerance weights are extracted from the exact probability mass of each discrete distance bin relative to the center, producing a parameter-free and scale-invariant metric.

For a 1-5 scale, the maximum possible distance is 4. For a 1-10 scale, the maximum distance is 9. Scaling the standard deviation proportionally by a factor of $9/4 = 2.25$, we model the expected inter-annotator variance on a 10-point scale as $\mathcal{N}(0, 2.25^2)$.

Applying the exact same Discrete Gaussian Derivation formulated in Section 3, we extract the probability mass for each integer distance d normalized against the central bin:

$$W(d) = \frac{\Phi(d + 0.5, \mu = 0, \sigma = 2.25) - \Phi(d - 0.5, \mu = 0, \sigma = 2.25)}{\Phi(0.5, \mu = 0, \sigma = 2.25) - \Phi(-0.5, \mu = 0, \sigma = 2.25)}$$

This derivation yields a dynamically scaled weight map, ensuring that distances represent the same qualitative variance regardless of the scale's resolution. As detailed in Table 7 and Figure 6, a distance of 2 points on a 1-10 scale receives a weight of 0.678, which maps almost accurately to the 0.631 weight assigned to a 1-point distance on a 1-5 scale.

Table 7: Discrete Gaussian weights for a 1-10 ordinal scale ($\sigma = 2.25$).

Distance (d)	**Tolerance Weight** (W)
0	1.000
1	**0.907**
2	0.678
3	0.417
4	0.211
≥ 5	≈ 0.000

C LLM-AS-A-JUDGE PROMPTS AND METHODOLOGY

To ensure full reproducibility of the In-Context Learning (ICL) explainability case study presented in this paper, we detail the core system prompts and evaluation methodology utilized for our LLM judge.

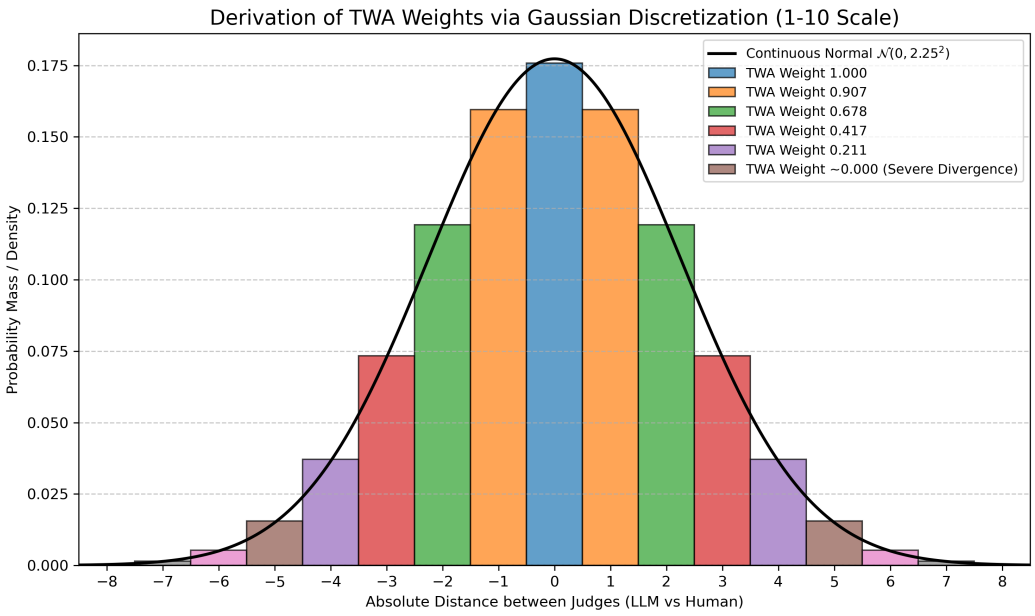


Figure 6: Discrete Gaussian Tolerance Weights for a 1-10 scale. The continuous distribution $\mathcal{N}(0, 2.25^2)$ is mapped to discrete point boundaries. large divergences ($d \geq 5$) receive practically zero weight, matching the behavior of the original 1-5 scale formulation.

C.1 System Prompt

The artificial judge was instructed using the following heavily anchored system prompt:

System Prompt: *You are an expert AI evaluator assessing the quality of explanations generated by a Multimodal Large Language Model (MLLM) under a few-shot in-context learning setup. The MLLM was given K examples of different classes and asked to classify a Query Image and explain its reasoning based ONLY on the visual evidence provided in the context.*

C.2 Condition Definitions

The judge evaluated the generated text based on four strictly defined explanatory conditions to ensure the evaluation was grounded in the requested format:

- **E2 (NLE):** Natural language explanation grounded only in observable visual evidence.
- **E3 (Features):** A minimal list of critical observable features formatted as bullet points.
- **E4 (Feature-value rules):** Rule-based explanations utilizing IF-THEN logic grounded in observable patterns.
- **E5 (DL Axioms):** Description Logic-based explanations utilizing formal TBox and ABox class conditions.

C.3 Detailed 1-5 Scoring Rubrics

To systematically quantify the quality of the explanations, the judge model (and the human annotators) were provided with the following detailed 1-5 rubrics for each dimension:

1. Textual Groundedness

Desideratum: Every relevant concept in the image is mentioned in the text explanation.
Definition (IMG $\rightarrow$ NLE): A well grounded vision-to-text explanation mentions every relevant

concept in the image in the explanation.
What it measures: The degree to which the generated text captures all salient objects, attributes, and spatial relationships present in the source image. Whether the text provides an exhaustive account of the visual scene without omitting relevant concepts.
Scores:

- **1 (Not Grounded):** large omissions of key concepts in the explanation.
- **2 (Weakly Grounded):** Major omissions of key concepts.
- **3 (Moderately Grounded):** Few omissions of key concepts.
- **4 (Strongly Grounded):** One omission of key concepts.
- **5 (accurately Grounded):** Zero omissions of key concepts.

2. HALLUCINATION-FREE

Desideratum: Every relevant concept in the explanation is visible in the image.
Definition (NLE $\rightarrow$ IMG): A hallucination-free explanation reflects accurate visual evidence.
What it measures: Evaluates whether the explanation reflects accurate visual evidence (any claims made in the explanation must be visually verifiable in the query image without hallucinating non-existent features).
Scores:

- **1:** large hallucinations
- **2:** Major hallucinations
- **3:** Few hallucinations
- **4:** One hallucination
- **5 (Hallucination-free):** Zero hallucinations.

3. CONCEPT COUNTING

Desideratum: The explanation accurately quantifies the exact number of relevant concepts present in the image.
Definition: A precisely counted explanation ensures that the elements counted in the NLE faithfully correspond with the number of concepts in the image.
What it measures: Evaluates the quantitative accuracy of the explanation (penalizing any discrepancies between the stated count of elements in the text and the actual number of countable elements visible in the image).
***Note*:** This metric is different from hallucination-free because it does reflect, for instance, a model explaining a flower of class f has 5 petals while the query image shows 6 petals.
Scores:

- **1 (Wrong concept counting):** 0% of the countable features are well counted.
- **2 (Weak concept counting):** 1–33% of the countable features are well counted.
- **3 (Moderate concept counting):** 34–66% of the countable features are well counted.
- **4 (Almost perfect concept counting):** 67–99% of the countable features are well counted.
- **5 (Perfect concept counting):** 100% of the countable features are well counted.

4. COMPREHENSIBILITY

Desideratum: The explanation is simple, readable, and fully accessible to an end-user without unnecessary complexity.
Definition: A highly comprehensible explanation is easy to read, simple, and accessible to end-users.
What it measures: Evaluates whether: 1) explanations are simple and accessible to end-users by measuring readability and penalizing overly complex or unnecessarily sophisticated explanations; 2) descriptions in terms of symbols are semantically and structurally similar to those of human experts.
Scores:

- **1 (Not Comprehensible):** Highly complex.
- **2 (Weak Comprehensibility):** Mostly complex.
- **3 (Medium Comprehensibility):** Somewhat complex.
- **4 (Strong Comprehensibility):** A light degree of complexity.
- **5 (accurately Comprehensible):** No complexity.

5. Conciseness

Desideratum: The explanation conveys only the strictly necessary information efficiently.
Definition: A concise explanation conveys the necessary information efficiently without unnecessary length nor fillers.
What it measures: Evaluates the length of the explanation (penalizing unnecessarily verbose answers, filler words, or repetition).
Scores:

- **1 (Not Concise):** Highly verbose, most content is redundant or removable.
- **2 (Weakly Concise):** Mostly verbose, several sentences can be removed without loss of meaning.
- **3 (Medium Conciseness):** Somewhat verbose, few sentences or parts can be removed while preserving full meaning.
- **4 (Strongly Concise):** Light verbosity, one sentence can be removed without losing meaning.
- **5 (accurately Concise):** No unnecessary verbosity, nothing can be removed without losing meaning.

6. Specificity

Desideratum: The explanation relies exclusively on precise, exact, and non-generic details about the sample being classified.
Definition: A highly specific explanation provides precise, concrete, and non-generic information about the classified subject. This metric is evaluated locally, meaning it only considers the individual sample at hand, without requiring knowledge of other classes.
What it measures: Evaluates whether the explanation relies on precise details on a local sample-basis (penalizing vague, broad, or generic statements).
Scores:

- **1 (Not Specific):** Highly vague, 0% of all features are precise, concrete, and non-generic.
- **2 (Weakly Specific):** Mostly vague, 1–33% of all features are precise, concrete, and non-generic.
- **3 (Moderately Specific):** Somewhat vague, 34–66% of all features are precise, concrete, and non-generic features.
- **4 (Strongly Specific):** Minimal vagueness, 67–99% of all features are precise, concrete, and non-generic.
- **5 (accurately Specific):** Zero vagueness, 100% of all features are precise, concrete, and non-generic.

8. Local Discriminativeness

Desideratum: The explanation successfully isolates the unique features that separate the predicted class from all others.
Definition: A discriminative explanation highlights the features and values that truly support the predicted class over all others, evaluated locally within the trial.
What it measures: Evaluates the capability of providing local key reasons causing a prediction.
Scores:

- **1 (Not Discriminative):** 0% pathognomonic features.

- **2 (Weakly Discriminative):** 1–33% pathognomonic features.
- **3 (Moderately Discriminative):** 34–66% pathognomonic features.
- **4 (Strongly Discriminative):** 67–99% pathognomonic features.
- **5 (accurately Discriminative):** 100% pathognomonic features.

9. INSTRUCTION FOLLOWING

Desideratum: The output flawlessly satisfies all explicit structure and format constraints.
Definition: An explanation that follows instructions is compliant with open-ended and format instructions.
What it measures: Evaluates how closely the model's output follows all explicit instructions and constraints in the prompt.
Scores:

- **1 (No Adherence):** 0% followed.
- **2 (Weak Adherence):** 1–33% followed.
- **3 (Partial Adherence):** 34–66% followed.
- **4 (Strong Adherence):** 67–99% followed.
- **5 (Perfect Adherence):** 100% followed.

10. LOGICAL COHERENCE

Desideratum: The sentences connect logically to form a valid, smooth, and complete deduction.
Definition: A logically coherent explanation builds from sentence to sentence into a valid body of information.
What it measures: Evaluates whether the explanation forms a valid logical progression.
Scores:

- **1 (Not Coherent):** Completely disconnected clauses.
- **2 (Weakly Coherent):** large logical errors.
- **3 (Moderately Coherent):** Some logical issues.
- **4 (Strongly Coherent):** Mostly smooth, minor error.
- **5 (accurately Coherent):** Flawless deduction.

D ADDITIONAL THEORETICAL CONSIDERATIONS

This appendix expands upon the mathematical and methodological foundations that support the necessity of Tolerance-Weighted Agreement (TWA) in modern LLM-as-a-judge pipelines.

D.1 EVALUATING CONSENSUS ACROSS THREE DIMENSIONS

Because humans naturally possess varying subjective thresholds (e.g., one expert's "good" is a 4, while another's is a 5), demanding rigid agreement on granular scales often masks true alignment. Therefore, evaluative alignment must be analyzed across three distinct dimensions. The first is *Human-Human Alignment* (the Inter-Annotator Baseline), which measures the natural subjective variance among human domain experts, acknowledging that humans rarely achieve perfect exact-match agreement on granular scales. The second is *LLM-LLM Alignment* (Inter-Model Alignment), which measures the congruence between different artificial judges, evaluating whether two distinct models consistently apply a shared rubric with the same semantic intent. Finally, *Hybrid Human-LLM Alignment* measures the degree to which an artificial judge successfully mimics the subjective scaling and evaluative intent of human consensus.

Tolerance-Weighted Agreement (TWA, ω) provides a framework that can be applied to these dimensions. By absorbing natural subjective variance, it offers a pragmatic alternative for assessing Inter-Annotator Agreement (Human-Human), cross-model congruence (LLM-LLM), and hybrid evaluator congruence (Human-LLM).

D.2 The Importance of Highly Granular Rubrics

The concept of natural subjective variance (where a human and an LLM might legitimately differ by 1 point on a Likert scale) relies on a fundamental assumption: the evaluation must be based on an ordinal discrete scale supported by unambiguous rubrics, allowing both judges to consistently interpret the magnitude of each score. The interpretability of the $d \leq 1$ tolerance window relies heavily on this semantic consistency. If a 1-5 scale is provided but the qualitative differences between intermediate steps (e.g., 2, 3, and 4) are ambiguous, human and artificial judges will naturally diverge. In such cases, TWA mathematically functions exactly as intended: it will correctly penalize this arbitrary variance and report a lack of alignment.Therefore, while TWA mathematically evaluates consensus regardless of the rubric's quality, successfully achieving Human-LLM alignment dictates that researchers support their ordinal scales with highly granular, well-defined scoring rubrics (such as those detailed in Appendix C). Each point on the scale must represent a distinct, qualitative evaluative step to ensure that minor deviations reflect genuine subjective calibration rather than random confusion.

D.3 The Argument Against Rank Correlations in Evaluation

While rank correlation coefficients (such as Spearman's ρ and Kendall's τ) are frequently used as alternatives to agreement metrics, they present three critical shortcomings in LLM-as-a-judge pipelines:

1. **Non-linear Interpretability:** A Spearman correlation of $\rho = 0.70$ does not mean the LLM is 70% aligned with the human, nor does it provide a threshold for when a researcher should trust the judge. It is an abstract statistical measure of monotonicity.
2. **Inability to Model Local subjective variance:** Rank correlations evaluate the entire global distribution simultaneously. They cannot answer the operational question: *"In what percentage of cases does the LLM provide a reliable, human-equivalent score?"* TWA answers this directly by thresholding local point deviations.
3. **Instability Due to High Tie Density:** Modern state-of-the-art LLMs generate high-quality text, causing evaluation distributions to heavily skew towards scores of 4 and 5. This high density of identical scores (ties) severely destabilizes the mathematical ranking formulas, yielding noisy and unintuitive metric readouts.

D.4 TWA vs. Krippendorff's Interval Alpha

In the SummEval dataset (Fabbri et al., 2021), the authors recognized the flaws of nominal metrics on 1-5 scales and utilized Krippendorff's Interval α to account for the distance between disagreements. While this provided massive theoretical justification for distance-aware evaluations, Krippendorff's Alpha still suffers from the Kappa Paradox.

Like Cohen's Kappa, Krippendorff's α is adjusted for chance agreement (p_e). Because modern datasets are highly skewed toward excellent outputs (lots of 4s and 5s), the metric mathematically assumes that the probability of chance agreement is extraordinarily high. Consequently, the coefficient artificially crashes, reporting a "poor alignment" even when the judge and the human agree 90% of the time. TWA solves this by abandoning the flawed chance-agreement penalty, measuring the absolute and semantically localized alignment robust to distribution skews.

E Annotator Stability Distribution

To further validate the robustness of the LLM-as-a-judge alignment against natural human subjectivity, Figure 7 maps the specific TWA alignment levels of each independent annotator across every individual evaluation dimension. The majority of cells fall in the "Aligned" (Green) or "Partially Aligned" (Yellow) categories, confirming that the LLM successfully models the natural subjective baseline of nearly every human judge.

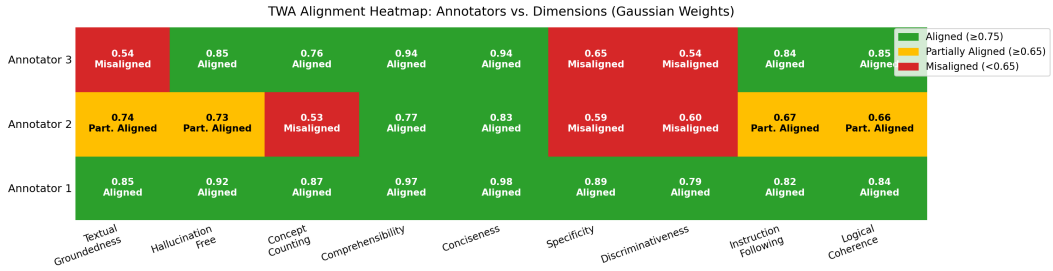


Figure 7: Heatmap of TWA alignment levels mapping Annotators vs. Dimensions. The color coding strictly reflects the operational alignment thresholds (Green ≥ 0.75, Yellow ≥ 0.65, Red < 0.65).

F Empirical Error Distributions and the Kappa Paradox

To validate the central premise of TWA—that human and LLM evaluator disagreements follow a normal distribution rather than uniform or random noise—we analyzed the empirical error distributions ($d = \text{LLM} - \text{Human}$) for the ICL explainability evaluation dataset and, separately, for the two state-of-the-art evaluation benchmarks (BiGGen-Bench and SummEval). We deliberately keep these two analyses in separate figures, generated independently from each dataset's own notebook: Figure 8 never mixes raw ICL scores with BiGGen-Bench or SummEval scores, and vice versa.

As shown in Figure 8, broken down into its Overall pooled distribution (across all 9 dimensions) plus three representative dimensions spanning the alignment range (Conciseness: highest TWA; Specificity: mid-range; Discriminativeness: lowest TWA), the Custom ICL error distributions are consistently unimodal, approximately bell-shaped, and centered close to $d = 0$ ($\mu \in [-0.32, 0.17]$), aligning closely with the standard normal assumption $\mathcal{N}(0, 1^2)$ embedded in the TWA discretization for 1-5 scales.

Figure 9 shows the same analysis for BiGGen-Bench and SummEval, each broken down into its Overall pooled distribution plus its highest- and lowest-TWA dimensions (BiGGen-Bench: Grounding, Refinement; SummEval: Consistency, Fluency). All three BiGGen-Bench panels are centered close to $d = 0$ ($\mu \in [0.01, 0.71]$). SummEval tells a different story: its Overall panel is already centered at $d \approx -1.12$, and every individual SummEval dimension we inspect is shifted negative, from Consistency ($d \approx -1.02$) to Fluency ($d \approx -2.04$). This indicates that the G-Eval/GPT-4 judge on SummEval carries a systematic under-scoring bias across the benchmark, not an isolated quirk of one dimension.

This pattern shows that TWA's discrete-Gaussian tolerance weights are a reasonable model for the *shape* of evaluator disagreement across these benchmarks, but that shape alone does not guarantee zero-centered agreement: a judge can be reliably bell-shaped in its errors while still being systematically miscalibrated, as G-Eval/GPT-4 on SummEval illustrates. We view detecting and correcting such offsets (e.g., via per-dimension mean-centering before applying TWA) as useful complementary diagnostics, and discuss this further in Section 5.

At the same time, these panels reaffirm why TWA's Misaligned label is the correct call for a case like SummEval Fluency, rather than an artifact of the metric being too strict. TWA is designed to extend meaningful credit only when judge and human scores coincide exactly or differ by at most one point; it is not designed to forgive larger, systematic gaps. In the Fluency panel, 88.5% of instances differ by two points or more ($d \leq -2$), well outside that tolerance window, and TWA's low score ($\omega = 0.196$) tracks this directly. This is the behavior TWA is meant to have: it distinguishes near-miss subjective noise, which it forgives, from large and consistent score gaps, which it flags. That distinction is what makes TWA useful in practice, not just as a diagnostic summary but as an actionable signal for iterating between reviewers and LLM judges to close the gap and reach better alignment.

Furthermore, to empirically demonstrate the failure of traditional chance-adjusted metrics under these skewed Gaussian distributions, we visualize the score differences for the *Conciseness* dimension from the ICL explainability evaluation dataset.

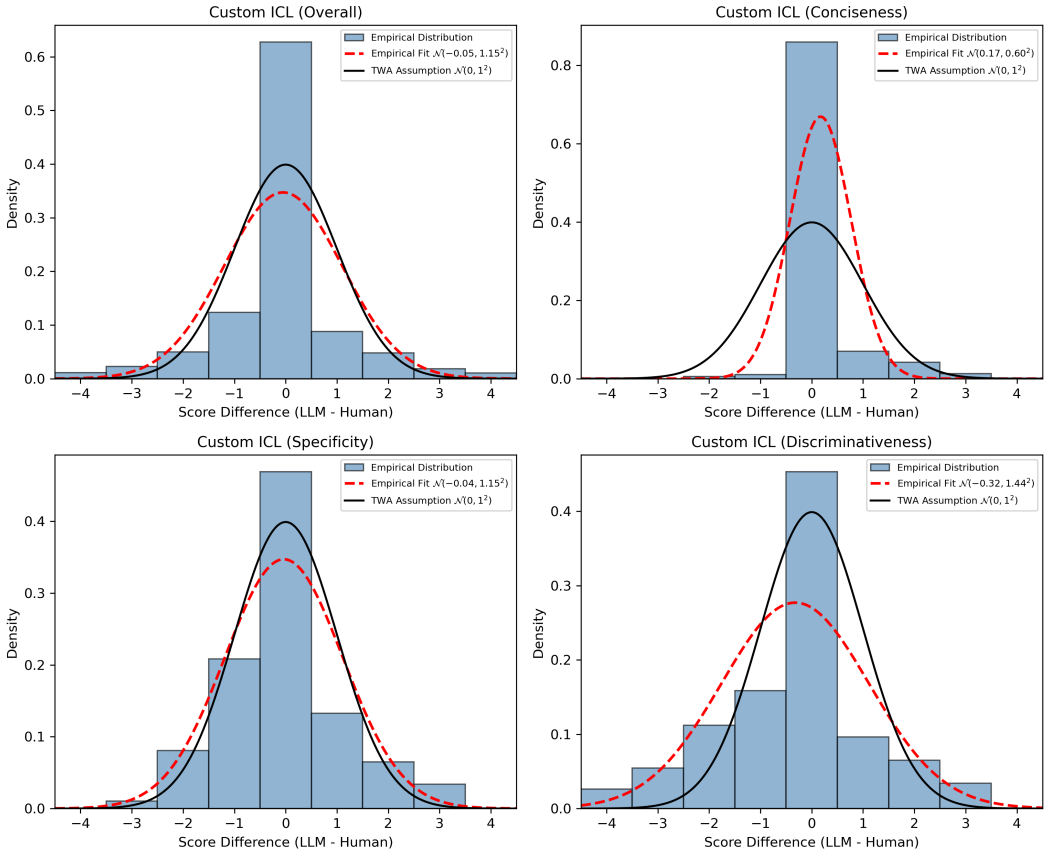


Figure 8: Empirical error distributions (d = LLM − Human) for the custom ICL explainability evaluation dataset: Overall plus three representative dimensions (Conciseness, Specificity, Discriminativeness). All four panels are unimodal, approximately bell-shaped, and centered close to $d = 0$, echoing the discrete Normal assumption of the TWA formulation (black solid line).

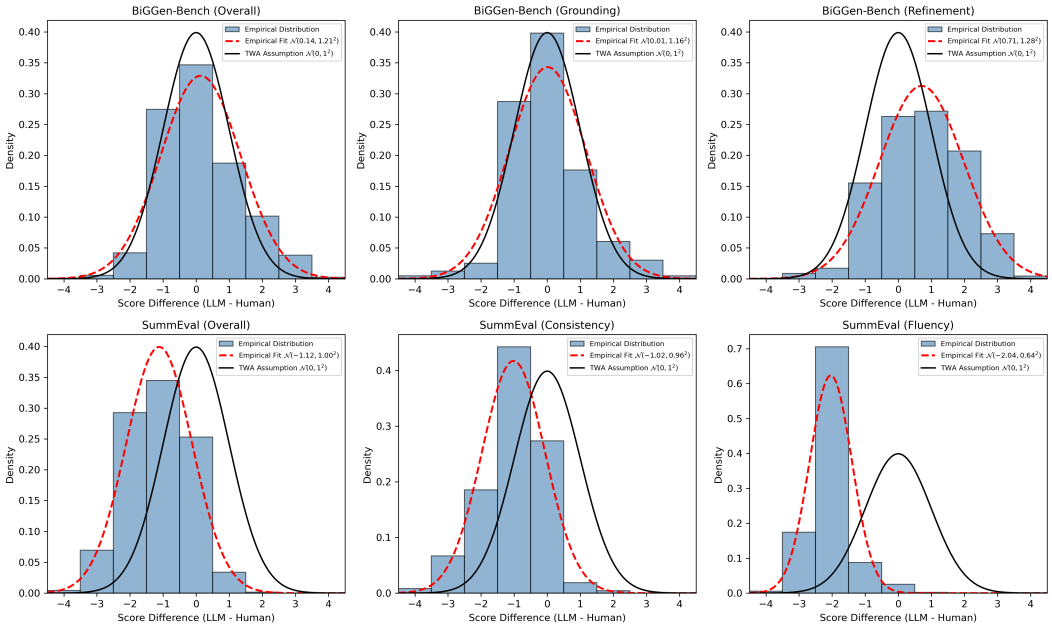


Figure 9: Empirical error distributions ($d = \text{LLM} - \text{Human}$) for BiGGen-Bench and SummEval: Overall plus each dataset's highest- and lowest-TWA dimension. BiGGen-Bench panels stay centered near $d \approx 0$; every SummEval panel is shifted negative (Overall $d \approx -1.12$, Fluency $d \approx -2.04$), reflecting a systematic judge under-scoring bias rather than symmetric subjective noise.

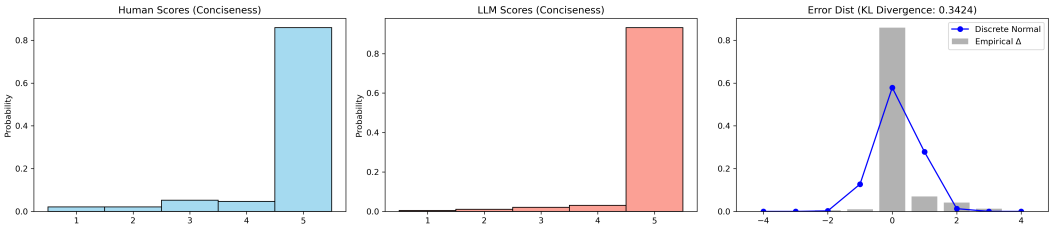


Figure 10: Comparison of the empirical error distribution against theoretical chance-adjusted models. While the LLM achieves nearly 86% Exact Match and 0.92 TWA, Cohen's Kappa collapses to 0.28 due to the extreme skewness of the underlying distribution. The low KL Divergence proves that the variance is natural Gaussian noise, not random evaluator hallucination.

This mathematical collapse is the direct result of the Kappa Paradox (Feinstein & Cicchetti, 1990), which we refer to throughout this work as *Kappa Deflation* in the context of LLM evaluation. When marginal rating distributions are heavily skewed (e.g., both human and LLM predominantly score outputs as 4 or 5), the expected chance agreement (p_e) in the formula $\kappa = \frac{p_o - p_e}{1 - p_e}$ artificially approaches 1.0. Consequently, even when the observed absolute agreement (p_o) is highly high (as seen with the 86% Exact Match in Figure 10), the numerator approaches zero, causing the κ coefficient to decrease. TWA bypasses this deflation fully by operating as an absolute, non-chance-adjusted metric rooted in Gaussian tolerance rather than theoretical random guessing.

G SOTA Dataset Error Magnitude Distribution

To complement the results presented in Table 6, we visualize the distribution of exact matches and partial tolerance deviations ($+1$ and -1 differences) for the evaluated state-of-the-art datasets (BiGGen-Bench and SummEval).

As shown in Figure 11, the majority of divergences across all external datasets fall within the $d \leq 1$ margin, reinforcing the theoretical premise of TWA. Even in datasets with lower Exact Match rates, large errors ($d \geq 2$) remain a statistical minority.

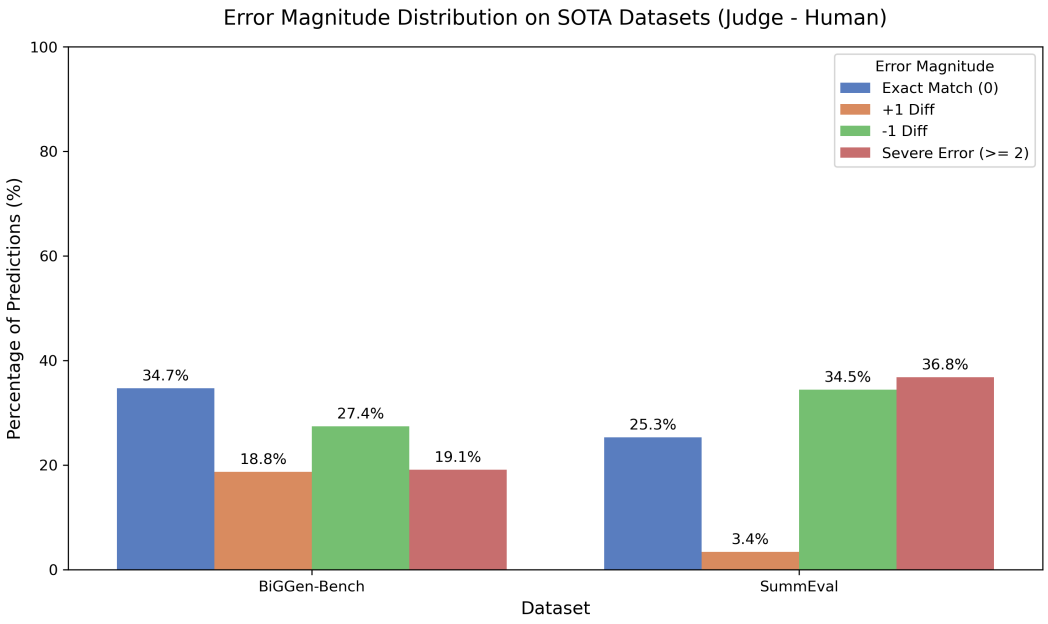


Figure 11: Distribution of Error Magnitudes across BiGGen-Bench and SummEval. The chart displays the percentage of Exact Matches (0), $+1$ differences, -1 differences, and large errors (≥ 2).

Table 8 gives the same breakdown at the per-dimension level. It makes the directional asymmetry behind Figure 9 explicit: on every SummEval dimension, -1 differences (judge scores below human) substantially outnumber $+1$ differences (e.g., Consistency: 44.2% vs. 1.9%; Relevance: 43.6% vs. 4.6%), whereas BiGGen-Bench's $+1/-1$ splits are comparatively balanced. This directional imbalance, not just the spread, is what drives SummEval's negative mean offset.

Table 8: Per-dimension error magnitude and tolerance breakdown for BiGGen-Bench and SummEval, complementing Figure 11. Rows are sorted by TWA within each dataset. The $+1$ and -1 differences highlight directional biases not captured by pure magnitude $|d|$, while $|d| \geq 2$ represents major errors. Aligned scores (TWA ≥ 0.75) would appear in **bold**; none do.

Dataset	Dimension	Exact Match	$+1$ Diff	-1 Diff	$\lvert d\rvert \geq 2$	MAE	TWA (Alignment)
BiGGen-Bench (Kim et al., 2025)	Grounding	39.8%	17.6%	28.7%	13.9%	0.80	0.705 (Partially Aligned)
BiGGen-Bench (Kim et al., 2025)	Safety	41.1%	21.8%	20.6%	16.5%	0.81	0.698 (Partially Aligned)
BiGGen-Bench (Kim et al., 2025)	Planning	38.2%	23.6%	23.2%	15.0%	0.79	0.698 (Partially Aligned)
BiGGen-Bench (Kim et al., 2025)	Theory Of Mind	35.2%	11.2%	35.8%	17.8%	0.86	0.672 (Partially Aligned)
BiGGen-Bench (Kim et al., 2025)	Reasoning	35.0%	15.8%	28.2%	21.0%	0.90	0.655 (Partially Aligned)
BiGGen-Bench (Kim et al., 2025)	Instruction Following	28.5%	19.0%	34.0%	18.5%	0.93	0.645 (Misaligned)
BiGGen-Bench (Kim et al., 2025)	Tool Usage	31.2%	19.7%	25.1%	24.0%	1.02	0.620 (Misaligned)
BiGGen-Bench (Kim et al., 2025)	Refinement	26.3%	27.2%	15.5%	31.0%	1.14	0.569 (Misaligned)
SummEval (Fabbri et al., 2021)	Relevance	36.7%	4.6%	43.6%	15.1%	0.79	0.694 (Partially Aligned)
SummEval (Fabbri et al., 2021)	Coherence	34.8%	6.9%	41.3%	17.0%	0.86	0.674 (Partially Aligned)
SummEval (Fabbri et al., 2021)	Consistency	27.4%	1.9%	44.2%	26.5%	1.07	0.596 (Misaligned)
SummEval (Fabbri et al., 2021)	Fluency	2.5%	0.2%	8.8%	88.5%	2.05	0.196 (Misaligned)

As a complementary summary view, Figure 12 averages Exact Match and TWA across dimensions for each dataset. It visualizes the pattern discussed in Section 4.4: only the Custom ICL dataset's

average crosses the Aligned threshold, while BiGGen-Bench and SummEval, though substantially higher under TWA than under Exact Match, remain below it on average.

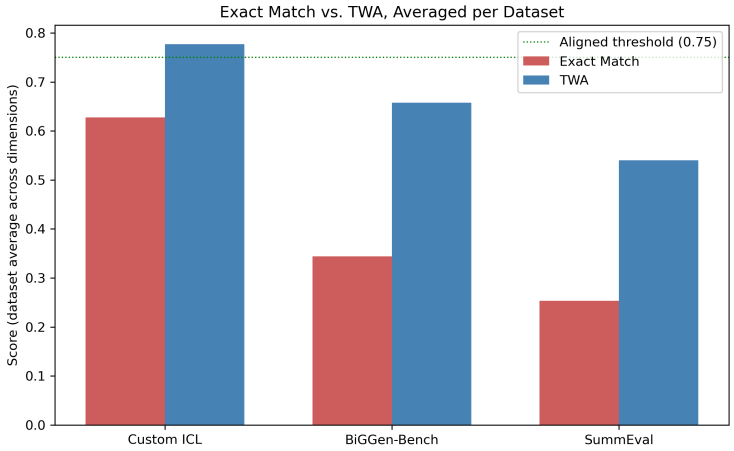


Figure 12: Exact Match vs. TWA, averaged across dimensions for each dataset. The dotted line marks the Aligned threshold ($\omega \geq 0.75$).

H Numerical Simulation Methodology

To systematically compare the behavior of Tolerance-Weighted Agreement (TWA) and Gwet's AC2 (using quadratic weights) under varying statistical conditions, we simulated 10,000 independent trial pairs for a ground truth and an algorithmic judge on a discrete 1-5 scale. For each experiment, we show both the empirical error distribution (left) and the resulting metric sensitivity curves (right).

Experiment A: Variance Impact We fixed the ground truth centered at $\mu = 3.0$ and increased the standard deviation of the injected noise ($\sigma_{error} \in [0.1, 3.0]$). As shown in Figure 13, both metrics degrade predictably as the variance increases, though TWA drops more sharply once errors exceed the $d = 1$ tolerance window.

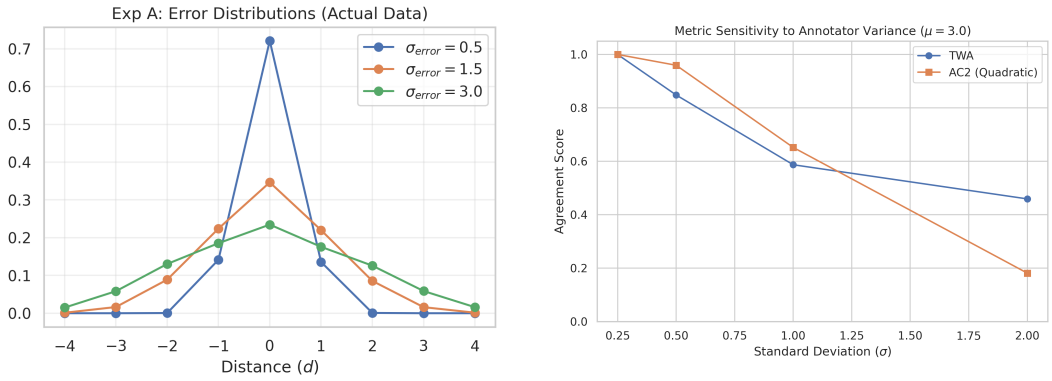


Figure 13: **Variance Impact.** Left: Subjective noise distributions $d = a_2 - a_1$ at varying standard deviations. Right: Both TWA and AC2 degrade as variance increases.

Experiment B: Sensitivity to Systematic Bias (Shifted Error) We fixed the noise standard deviation at $\sigma_{error} = 1.0$, but shifted the mean of the error (μ_{error}) from -3.0 to +3.0 to simulate a systematically biased judge (overly harsh or overly lenient). As shown in Figure 14, both metrics symmetrically and heavily penalize systematic shifts, forming an inverted-V. This demonstrates that TWA rigorously drops when a model systematically hallucinates in one direction, independent of the ground truth marginals.

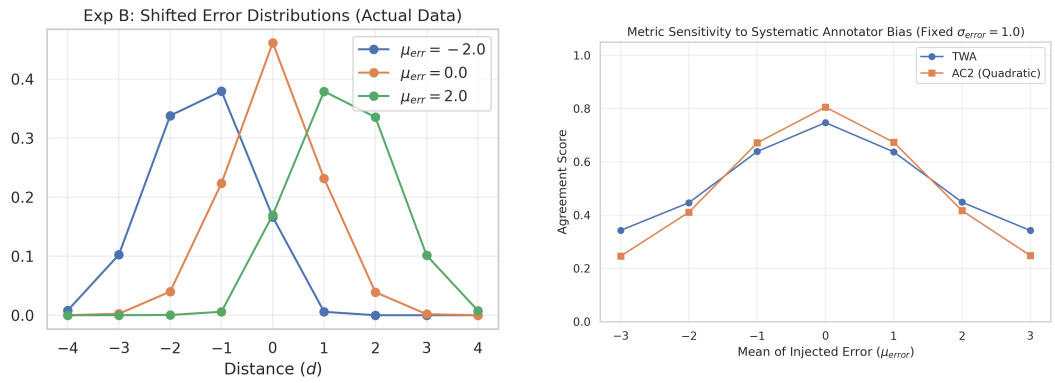


Figure 14: **Systematic Bias.** Left: Error distributions shifted left (harsh judge) and right (lenient judge). Right: Both metrics heavily penalize systematic mean shifts.

Experiment C: Sensitivity to Kurtosis (Heavy Tails) To isolate the effect of large, rare errors ($|d| \geq 2$) from general variance, we employed a Discrete Generalized Normal distribution, sweeping the shape parameter β. For every β, we mathematically constrained the overall discrete variance to be strictly $\sigma^2 = 1.0$. Thus, higher kurtosis means fewer small errors but more extreme outliers. As shown in Figure 15, TWA is highly sensitive to heavy tails and plunges rapidly. AC2, due to its quadratic formulation, is far more forgiving of large outliers and remains artificially inflated.

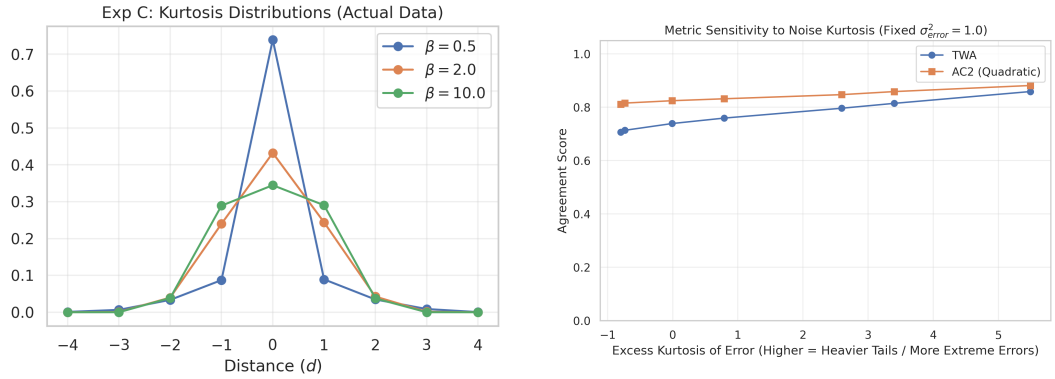


Figure 15: **Kurtosis Impact.** Left: Error distributions with varying tail weights (fixed variance $\sigma^2 = 1.0$). Right: TWA severely penalizes the extreme outliers (heavy tails), whereas AC2 remains forgiving due to its quadratic partial credit.

I Statistical Significance and Permutation Testing

A common criticism of absolute evaluation metrics that do not mathematically correct for chance agreement is that their reported scores might simply reflect two independent judges randomly guessing according to their marginal distributions. To rigorously verify that Tolerance-Weighted Agreement (TWA) accurately distinguishes true evaluative consensus from random noise, we conducted a non-parametric permutation test ($K = 10,000$ iterations) across all datasets analyzed in this study.

For each evaluated pair of judges (Human vs. LLM, or Human vs. Human), we computed the observed TWA, ω_{obs}. We then generated a null distribution by shuffling the judge scores 10,000 times while keeping the target scores fixed. This destroys any true item-level correspondence while exactly preserving the empirical marginal distributions of both judges. The p-value was calculated as the fraction of permutations where $\omega_{null} \geq \omega_{obs}$. To rigorously control for Type I errors across the $m = 43$ simultaneous hypothesis tests, we applied the Benjamini-Hochberg (FDR) correction to compute adjusted p-values (p_{adj}).

As detailed in Table 9, across the Custom ICL dataset and BiGGen-Bench, the LLM judges were rigorously distinguishable from chance ($p_{adj} < 0.05$, $z > 2$). Crucially, in the SummEval benchmark, the G-Eval model's performance on the Fluency dimension—which TWA explicitly flagged as Misaligned ($\omega = 0.196$)—yielded a z-score of -7.04 ($p \approx 1.0000$ for the right-tailed test), indicating a significant discordance: the model performs systematically worse than a judge rolling dice based on its marginal biases, driven by its severe under-scoring. Furthermore, in the human Inter-Annotator Agreement (IAA) matrix, some annotator pairs also exhibited indistinguishable-from-chance agreement even on nominally objective dimensions like *Concept Counting*, highlighting true evaluative noise or annotator fatigue where a 1-point deviation represents a factual error. These results confirm that TWA acts as a rigorous statistical filter for large misalignment and pure noise, effectively eliminating the need for complex chance-adjusted covariance matrices like Cohen's Kappa or Krippendorff's Alpha.

Table 9: Permutation test results ($K = 10,000$) across all evaluation pairs. The adjusted p-value (p_{adj}) applies Benjamini-Hochberg (FDR) correction for multiple comparisons ($m = 43$). Dimensions indistinguishable from chance ($p_{adj} \geq 0.05$) or showing significant discordance ($z \ll 0$) are marked accordingly.

Dataset	**Dimension**	N	ω_{obs}	μ_{null}	p_{adj}	z-score	**Verdict**
ICL Explainability	Textual Groundedness	384	0.711	0.472	0.0000	15.14	Distinguishable
ICL Explainability	Hallucination Free	384	0.834	0.768	0.0000	7.74	Distinguishable
ICL Explainability	Concept Counting	384	0.720	0.670	0.0000	4.54	Distinguishable
ICL Explainability	Comprehensibility	384	0.894	0.785	0.0000	15.12	Distinguishable
ICL Explainability	Conciseness	384	0.918	0.862	0.0000	11.35	Distinguishable
ICL Explainability	Specificity	384	0.708	0.491	0.0000	13.98	Distinguishable
ICL Explainability	Discriminativeness	384	0.644	0.441	0.0000	10.58	Distinguishable
ICL Explainability	Instruction Following	384	0.779	0.581	0.0000	12.69	Distinguishable
ICL Explainability	Logical Coherence	384	0.787	0.614	0.0000	11.94	Distinguishable
ICL IAA (Pair 1)	Textual Groundedness	64	0.796	0.454	0.0000	7.50	Distinguishable
ICL IAA (Pair 1)	Hallucination Free	64	0.668	0.615	0.0097	2.63	Distinguishable
ICL IAA (Pair 1)	Concept Counting	64	0.430	0.407	0.0308	1.71	Distinguishable
ICL IAA (Pair 1)	Comprehensibility	64	0.656	0.527	0.0000	4.41	Distinguishable
ICL IAA (Pair 1)	Conciseness	64	0.923	0.841	0.0000	5.90	Distinguishable
ICL IAA (Pair 1)	Specificity	64	0.615	0.413	0.0000	5.17	Distinguishable
ICL IAA (Pair 1)	Discriminativeness	64	0.658	0.397	0.0000	5.48	Distinguishable
ICL IAA (Pair 1)	Instruction Following	64	0.642	0.444	0.0000	4.21	Distinguishable
ICL IAA (Pair 1)	Logical Coherence	64	0.587	0.474	0.0099	2.50	Distinguishable
ICL IAA (Pair 2)	Textual Groundedness	64	0.637	0.575	0.0133	2.30	Distinguishable
ICL IAA (Pair 2)	Hallucination Free	64	0.888	0.839	0.0061	3.38	Distinguishable
ICL IAA (Pair 2)	Concept Counting	64	0.837	0.816	0.1374	1.46	Indistinguishable
ICL IAA (Pair 2)	Comprehensibility	64	0.964	0.925	0.0045	5.39	Distinguishable
ICL IAA (Pair 2)	Conciseness	64	0.941	0.897	0.0007	5.48	Distinguishable
ICL IAA (Pair 2)	Specificity	64	0.748	0.658	0.0029	3.16	Distinguishable
ICL IAA (Pair 2)	Discriminativeness	64	0.696	0.591	0.0017	3.13	Distinguishable
ICL IAA (Pair 2)	Instruction Following	64	0.866	0.859	0.1661	0.72	Indistinguishable
ICL IAA (Pair 2)	Logical Coherence	64	0.815	0.662	0.0000	5.17	Distinguishable
ICL IAA (Pair 3)	Textual Groundedness	64	0.561	0.504	0.0247	1.99	Distinguishable
ICL IAA (Pair 3)	Hallucination Free	64	0.889	0.791	0.0000	5.24	Distinguishable
ICL IAA (Pair 3)	Concept Counting	64	0.622	0.592	0.1456	1.12	Indistinguishable
ICL IAA (Pair 3)	Comprehensibility	64	0.900	0.866	0.0211	2.64	Distinguishable
ICL IAA (Pair 3)	Conciseness	64	0.789	0.707	0.0005	3.72	Distinguishable
ICL IAA (Pair 3)	Specificity	64	0.453	0.445	0.3739	0.37	Indistinguishable
ICL IAA (Pair 3)	Discriminativeness	64	0.468	0.441	0.1873	0.93	Indistinguishable
ICL IAA (Pair 3)	Instruction Following	64	0.916	0.887	0.0215	2.79	Distinguishable
ICL IAA (Pair 3)	Logical Coherence	64	0.801	0.733	0.0079	2.71	Distinguishable
BiGGen (Inter-Judge)	Human vs Gpt4	2776	0.697	0.482	0.0000	35.41	Distinguishable
BiGGen (Inter-Judge)	Human vs Claude	2776	0.679	0.489	0.0000	32.14	Distinguishable
BiGGen (Inter-Judge)	Gpt4 vs Claude	2780	0.798	0.537	0.0000	47.27	Distinguishable
SummEval	Coherence	1600	0.674	0.555	0.0000	19.97	Distinguishable
SummEval	Consistency	1600	0.596	0.508	0.0000	18.18	Distinguishable
SummEval	Fluency	1600	0.196	0.209	1.0000	-7.04	Sig. Discordance
SummEval	Relevance	1600	0.694	0.598	0.0000	17.76	Distinguishable

J GWET'S AC2 FORMULA FOR ORDINAL AGREEMENT

While Section 2.3 highlights the structural limitations of Gwet's AC2 for real-time LLM evaluation, we provide the complete mathematical formulation (Gwet, 2014) here for completeness.

Given a dataset with q ordinal categories (e.g., $q = 5$ for a 1-5 scale), the AC2 statistic ($\widehat{\gamma}_2$) utilizes a weighted agreement formula:

$$\hat{\gamma}_2 = \frac{p_a - p_e}{1 - p_e}, \text{ where } \begin{cases} p_a & = \sum_{k,l}^{q} w_{kl} p'_{kl}, \\ p_e & = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi_k (1 - \pi_k). \end{cases} \tag{3}$$

The various terms in equation 5.6.1 are defined as follows:

- q is the number of scores used in the experiment,
- w_{kl} is the weight associated with the two categories k and l,
- $p'_{kl} = p_{kl}/\theta$, where p_{kl} is the relative number of subjects that raters 1 and 2 scored as x_k and x_l respectively, and θ the relative number of subjects scored by both raters (i.e. subjects with no missing rating).
- T_w is the sum of all weights w_{kl} associated with all categories.
- π_k is the probability that a rater assigns score x_k to a subject, and is calculated as $\pi_k = (p'_{k+} + p'_{+k})/2$. Note that $p'_{k+} = p_{k+}/\theta_1$ with θ_1 being the relative number of subjects that rater 1 has scored, and p_{k+} the relative number of subjects that rater 1 scored as x_k. Likewise, $p'_{+k} = p_{+k}/\theta_2$ with θ_2 being the relative number of subjects that rater 2 has scored, and p_{+k} the relative number of subjects that rater 2 scored as x_k.

This exact formulation demonstrates why AC2 cannot be calculated point-by-point. The P_e^w term strictly couples the penalty for any individual error to the global distribution (π_k) of the entire evaluation batch.

K MATHEMATICAL DERIVATION OF TWA THRESHOLDS

To complement the operational thresholds established in Section 4.4, we provide the mathematical derivations used to anchor the *Partially Aligned* ($\omega = 0.65$) and *Aligned* ($\omega = 0.75$) boundaries. These thresholds are not arbitrary; they are strictly derived from deterministic error profiles that represent baseline behaviors for an artificial judge.

1. The Random Baseline: First, we must ensure our thresholds mathematically exceed the expected value of a uniformly random judge. For a discrete 1-5 scale, a random evaluator assigning scores $x \in \{1, 2, 3, 4, 5\}$ with uniform probability $P(x) = 0.2$ against a target distribution also distributed uniformly will yield the following distance probabilities:

- $P(|d| = 0) = \frac{5}{25} = 20\%$
- $P(|d| = 1) = \frac{8}{25} = 32\%$
- $P(|d| = 2) = \frac{6}{25} = 24\%$
- $P(|d| = 3) = \frac{4}{25} = 16\%$
- $P(|d| = 4) = \frac{2}{25} = 8\%$

Applying the Discrete Gaussian tolerance weights $W(|d|)$:

$$E[\omega_{random}] = (0.20 \times 1.0) + (0.32 \times 0.631) + (0.24 \times 0.158) + 0 + 0 = 0.4398$$

Thus, a purely random judge achieves $\omega \approx 0.44$. The Partially Aligned threshold (0.65) is structurally isolated from this noise floor.

The Skewed Baseline: While the uniform baseline establishes the absolute mathematical floor, modern LLMs frequently exhibit highly skewed score distributions (e.g., generating mostly 4s and 5s). To verify TWA's robustness against "lazy" skewed guessing (a core driver of the Kappa Paradox),

we compute the expected baseline for a judge that randomly guesses with a severe left skew (e.g., $P(1) = 0.05, P(2) = 0.05, P(3) = 0.1, P(4) = 0.3, P(5) = 0.5$) against a similarly skewed human distribution. Under these conditions, the expected baseline score rises to $E[\omega_{random_skewed}] \approx 0.614$. Because this value remains strictly below our Partially Aligned threshold (0.65), TWA proves exceptionally robust: even a highly skewed random judge cannot mathematically masquerade as an aligned evaluator.

2. Deterministic Edge-Case Error Profiles: To understand the operational meaning of these thresholds, we evaluate TWA under two deterministic edge-case error profiles (Figure 16): *Scenario A* and *Scenario B*.

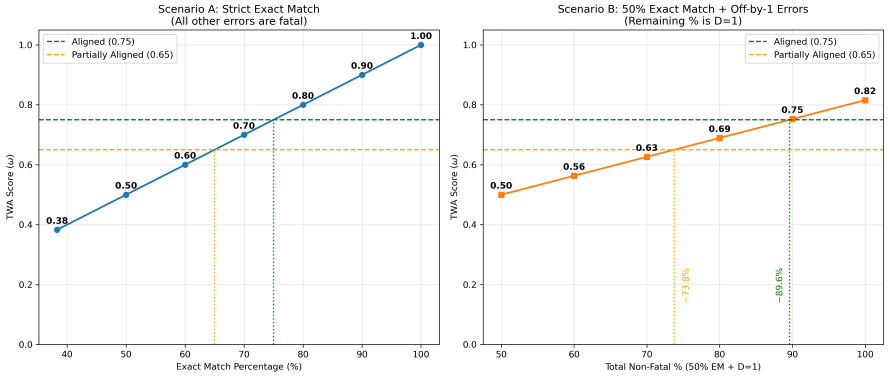


Figure 16: TWA Threshold Justification under Deterministic Error Profiles.

Scenario A: Pure Exact Match (Zero Tolerance) If a strict judge produces zero subjective variance ($P(|d| = 1) = 0$) and either exactly matches the target or fails substantially ($|d| \geq 3$), the TWA score reduces strictly to the Exact Match percentage. Under this rigorous constraint, reaching $\omega = 0.75$ requires the judge to be correct 75% of the time, and reaching $\omega = 0.65$ requires 65% perfection.

Scenario B: High Subjective Variance If a lenient judge only achieves exactly 50% Exact Match, how much subjective variance (off-by-1) is permissible before the judge is deemed misaligned? Using the formula: $\omega = P(|d| = 0) + P(|d| = 1) \times 0.631$.

- To reach $\omega = 0.65$: $0.65 = 0.50 + P(|d| = 1) \times 0.631 \implies P(|d| = 1) \approx 23.8\%$.
- To reach $\omega = 0.75$: $0.75 = 0.50 + P(|d| = 1) \times 0.631 \implies P(|d| = 1) \approx 39.6\%$.

Therefore, a judge anchored at 50% Exact Match can only be considered *Aligned* ($\omega = 0.75$) if its remaining errors are strictly adjacent 39.6% of the time, meaning a total of $\sim 90\%$ of all grading decisions must fall within a ≤ 1 point radius. This guarantees that an *Aligned* judge rarely fails severely.

3. The Implied Variance Bound ($\sigma < 1$): Finally, because the foundational tolerance weights $W(d)$ assume that natural subjective disagreement follows a standard normal distribution ($\sigma = 1.0$), an evaluator perfectly mimicking this theoretical baseline yields an expected TWA of $\omega \approx 0.707$. Establishing the Aligned threshold at $\omega \geq 0.75$ imposes a strict mathematical mandate: the artificial judge and the human ground truth must achieve a tighter consensus than the theoretical human baseline. Specifically, the distribution of their score differences must operate with an implied standard deviation of approximately $\sigma \leq 0.85$. This guarantees that an Aligned judge does not merely replicate standard subjective noise, but actively demonstrates a higher degree of evaluative convergence with the human reference.