ALTRA-TEST: AN ALTERNATIVE ANNOTATOR TEST FOR LLM JUDGES WITHOUT SHARED ANNOTATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as judges and graders to replace human annotators, raising the question of when such replacement is statistically justified. To support such decisions, the Alternative Annotator Test (alt-test) was introduced for dense settings where all items are annotated by the same annotators. However, in many large-scale annotation settings, items are typically annotated by different subsets of annotators. We therefore propose the Alternative Annotator Test with Resampled Annotations (ALTRA-TEST), a statistical test that does not require shared annotators across all items. ALTRA-TEST uses a bootstrap that resamples items, draws k human annotations for each item, and substitutes one of them with an LLM annotation. This produces paired all-human and mixed human–LLM panels. A replacement is supported when the inter-annotator agreement (IAA) of the mixed panel is statistically non-inferior to the IAA of the all-human panel. We first establish that ALTRA-TEST is a consistent test of asymptotic level α. We then evaluate it through finite-sample size simulations and real-world experiments in which we test 18 LLMs and 5 prompting techniques on 20 datasets spanning four modalities: text, image, audio, and video. Across these evaluations, ALTRA-TEST requires as few as 100 items to yield stable replacement decisions. In simulations, it dominates baselines by improving the trade-off: for any fixed rate of falsely accepting models noisier than simulated humans, , it correctly accepts more models that are less noisy. In real-world data, ALTRA-TEST supports replacing humans with LLMs in about half of the evaluated setups.

1 INTRODUCTION

Creating labeled datasets has long depended on human annotation, one of the slowest and most expensive stages of the AI pipeline (Ding et al., 2023; Tan et al., 2024). Researchers use human annotations not only to train models, but also as evaluation signals from which they draw empirical conclusions (Mei et al., 2026; Omar et al., 2026). Consequently, annotation quality directly affects the validity of the resulting models and findings. Large language models (LLMs) offer an appealing alternative: they can annotate quickly, cheaply, and at a scale that is difficult to achieve with experts or even with crowd workers (Chiang & Lee, 2023; Gilardi et al., 2023; Törnberg, 2023; Nahum et al., 2025). Indeed, LLMs are increasingly used as annotators, judges (LLM-as-a-judge), and raters (Zheng et al., 2023; Wei et al., 2024; Li et al., 2025; Wu, 2025; Mei et al., 2026).

Replacing human annotators with LLMs is useful only if the replacement preserves the relevant human signal (Egami et al., 2023; Gligoric et al., 2025; Wang et al., 2025). For objective tasks or those requiring domain expertise, the gold label is typically determined by expert judgment or aggregation (Rein et al., 2023; Nahum et al., 2025). However, LLMs have been shown to disagree with experts even when their responses appear fluent or confident (Zhou et al., 2024; Griot et al., 2025; Szymanski et al., 2025; Tseng et al., 2025; Wu & Aji, 2025). For more subjective tasks, such as preference annotation, this signal can be a distribution of human judgments, which have been shown to differ from LLM annotations (Lissak et al., 2024; Elangovan et al., 2025; Zhang et al., 2025). Therefore, using an LLM as an alternative to human annotators requires statistically justifying the replacement on the target annotation task (Bavaresco et al., 2025; Calderon et al., 2025b).

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.

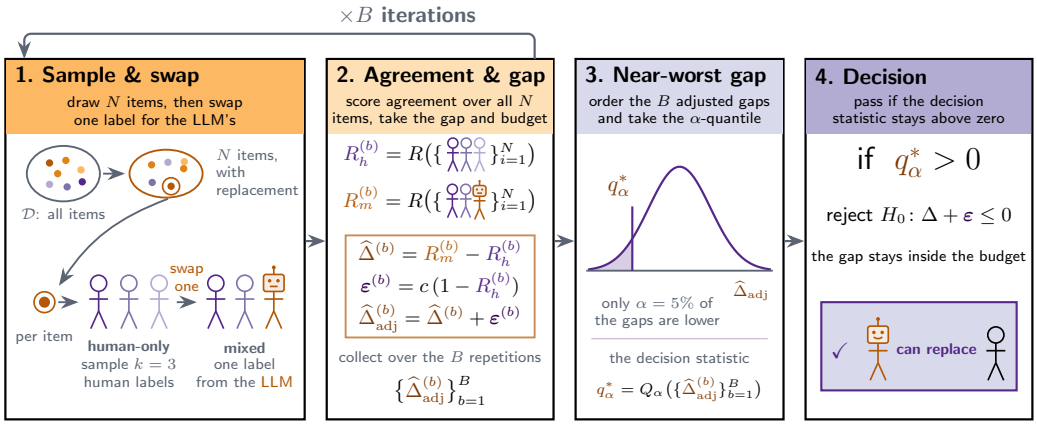


Figure 1: An illustration of the ALTRA-TEST. The two left boxes present the paired-bootstrap procedure: for each of B iterations, we sample N items and $k = 3$ annotations per item, replace one annotation with the LLM annotation, and compute IAA for the human-only and human–LLM panels. The substitution gap $\Delta^{(b)}$ and adaptive margin $\varepsilon^{(b)}$ are then computed. The two right boxes present the decision stage, based on a non-inferiority test. Here, q^*_α is the α-quantile statistic; if it is positive, we conclude that the LLM is a good alternative to a human annotator.

When ground-truth labels are unavailable, the correctness of human annotations can not be directly evaluated (Raykar et al., 2010; Paun et al., 2018). To assess the quality of the annotation process, researchers therefore use inter-annotator agreement (IAA) (Hayes & Krippendorff, 2007; Klie et al., 2024): if independent annotators consistently agree, the task is likely well-defined, the annotators share a common understanding of it, and the resulting labels capture a meaningful signal (Artstein & Poesio, 2008; Krippendorff, 2011). When evaluating whether an annotator LLM can serve as a substitute, the same standard should apply. However, researchers typically evaluate LLM annotations using accuracy, F_1, or correlation with a majority-vote label (Ziems et al., 2024; Bavaresco et al., 2025; Elangovan et al., 2025; Guerdan et al., 2025). These metrics quantify alignment but do not provide a statistically grounded replacement decision. Moreover, majority vote may collapse human variation into a single label, discarding disagreement that in many subjective tasks reflects legitimate perspectives (Basile et al., 2021; Davani et al., 2022; Plank, 2022; Sap et al., 2022; Fleisig et al., 2023). A replacement test should therefore mirror how we evaluate human annotation quality: through IAA that reflects reliability and respects all individual annotations.

Two recent works formulate human replacement by LLMs as a statistical procedure. Both assume a representative set of items, each labeled by multiple human annotators. Crucially, they also require the same group of annotators to label every item in this set. The Alternative Annotator Test (ALT-TEST; Calderon et al. (2025b)) holds out each annotator in turn and tests whether an LLM is non-inferior to that annotator in its agreement with the remaining panel. The Statistical Turing Test (STT; He et al. (2025a)) tests whether the IAA of human-only and mixed human–LLM panels is statistically equivalent. However, requiring a common annotator panel is restrictive in large-scale annotation settings. Annotators may label only small subsets of items, assignments may depend on expertise or availability, and datasets may omit annotator identities or links between annotators and labels (Siththaranjan et al., 2024). In these settings, researchers observe multiple human labels for each item but cannot identify a shared panel across items, making existing tests inapplicable.

In this work, we introduce the ***Alt**ernative annotator test with **R**esampled **A**nnotations* (ALTRA-TEST[1]; Figure 1), which does not require shared annotators across items. The test uses a paired percentile bootstrap: in each bootstrap sample, it resamples items with replacement, draws k human annotations per item, and constructs a mixed panel by replacing one human annotation with the LLM annotation. It then tests whether the mixed-panel IAA is non-inferior to the human-only IAA. An adaptive margin permits a small IAA loss proportional to the observed human disagreement. E.g., our calibrated default of $c = 0.1$ allows at most a 10% increase in disagreement. We say that an LLM *passes* the test when we reject the null hypothesis that the mixed-panel IAA is inferior.

[1] 🤗 anonymous-hf.com/a/6y60hp9vqg6k

We theoretically establish that ALTRA-TEST is a consistent test of asymptotic level α: as the number of items grows, the probability of correctly accepting a genuinely better model as an alternative to a human approaches 1, while the probability of falsely accepting an inferior model is strictly bounded by α. We then use simulations to study the test with practical sample sizes. We generate ground-truth labels and control whether the simulated LLM is more or less likely than simulated human annotators to recover them. This setup enables us to evaluate the replacement question directly, as we can measure each procedure's better-model pass rate and worse-model pass rate (i.e., correctly accepting simulated LLMs that are less noisy than simulated humans versus falsely accepting noisier ones). Across simulation settings, ALTRA-TEST dominates ALT-TEST and STT, and for the same worse-model pass rate, it passes more better models. The simulations also calibrate the default margin constant to $c = 0.1$ and show reliable behavior with around 100 items.

In addition, we evaluate the ALTRA-TEST on 20 real-world datasets across text, image, audio, and video, spanning objective and subjective tasks with a wide range of human IAA levels. We test 18 LLM judges and 5 prompting techniques. Under the default CoT setting, LLMs pass the test in 71 of 201 judge–dataset pairs (35.3%); when each judge uses its best prompting technique, this increases to 96 of 201 (47.8%). We find that replacement is strongly dataset-dependent, and in many passing cases the LLM not only passes the test but increases IAA. Consistent with the simulations, sample-size analyses show that around 100 items often suffice for strong judges to pass.

We encourage researchers to apply the ALTRA-TEST before delegating annotation to an LLM, and hope that it supports more explicit and statistically justified decisions.

2 RELATED WORK

IAA as a proxy for human annotation quality. When ground-truth labels are unavailable, researchers commonly assess annotation quality through IAA (Artstein & Poesio, 2008; Krippendorff, 2011). In this work, we use Krippendorff's α, a standard chance-corrected IAA coefficient that supports multiple annotators, missing annotations, and nominal or ordinal labels (Hayes & Krippendorff, 2007; Krippendorff, 2011). IAA scores should be interpreted relative to the task: objective or well-specified tasks are expected to yield high agreement, while difficult, ambiguous, or preference-based tasks may yield lower but still meaningful agreement (Pavlick & Kwiatkowski, 2019; Plank, 2022; Lissak et al., 2024). This interpretation depends on the purpose of annotation. In some settings, particularly in the social sciences, disagreement is not a defect but the object of study: researchers collect annotations precisely to characterize human perspectives and distributions of judgments (Aroyo & Welty, 2015; Basile et al., 2021; Uma et al., 2021; Plank, 2022). In such cases, forcing high IAA may remove the variation the study aims to measure. Our focus is different. We study AI annotation settings in which human labels are used as a measurement signal for training, evaluation, or judging models. Even in subjective tasks, such as response-preference annotation, where human IAA may have an inherent ceiling, positive and non-negligible IAA (e.g., above 0.2) indicates that the annotations carry a shared signal that an LLM replacement should reflect. Accordingly, we test whether LLM annotations preserve this signal.

Testing whether LLMs can replace human annotators. Various methods propose correcting or calibrating model-generated labels (Angelopoulos et al., 2023; Lee et al., 2025; Feng et al., 2026), but they do not decide whether replacement is justified. Two recent procedures address replacement as a statistical decision. The ALT-TEST (Calderon et al., 2025b) tests whether an LLM is non-inferior to a random held-out annotator in its agreement with the remaining annotators. The STT (He et al., 2025a) tests whether human and mixed human–LLM panels are statistically equivalent. Both require shared annotators across items, with each annotator labeling many items. This requirement is restrictive in many large-scale annotation setups, where annotators often label only small subsets of items, assignment is not fixed in advance, and worker participation is difficult to control. The ALTRA-TEST removes this requirement. Like ALT-TEST and unlike STT, it answers a non-inferiority question rather than an equivalence question. Unlike ALT-TEST and like STT, it uses a bootstrap and resampling of annotations. Unlike both, it does not replace all annotations from one annotator. Instead, for each sampled item, it draws k human annotations and replaces one of them with the LLM annotation and then bases the decision on a percentile-bootstrap statistic. We compare the ALTRA-TEST with ALT-TEST and STT and show that it dominates them for both non-inferiority and equivalence formulations.

3 METHOD

Setup and notations. Let $k \geq 3$ be the panel size used by the test, and let $\mathcal{D} = \{x_1, \ldots, x_N\}$ be a dataset of N items. For each item x_i, we observe M_i human annotations $H_i = \{h_{i,1}, \ldots, h_{i,M_i}\}$, with $M_i \geq k$. We do not require annotator identities, and we use each item's annotations only as a set of labels. The LLM annotation is $f(x_i)$. For each item, let $\mathbf{y}_{i,h}$ be a panel of k annotations from H_i, and let $\mathbf{y}_{i,m}$ be formed by replacing one annotation in $\mathbf{y}_{i,h}$ with $f(x_i)$. Stacking these item-level panels gives two annotation matrices: the human-only matrix $\mathbf{Y}_h$ and the mixed matrix $\mathbf{Y}_m$, both of dimension $N \times k$. Let R be an agreement or reliability score applied to an annotation matrix, where higher values indicate a more reliable annotation panel. We define $1-R$ as the disagreement level. In the main experiments, we use Krippendorff's α, the standard chance-corrected IAA coefficient for nominal and ordinal labels; we give its formal definition in Appendix B.1. Finally, we define:

$$R_h = R(\mathbf{Y}_h), \qquad R_m = R(\mathbf{Y}_m), \qquad \Delta = R_m - R_h.$$

The substitution gap Δ measures the effect of replacement; a positive value means that the mixed panel is more reliable than the human-only panel.

3.1 THE ALTRA-TEST

The ALTRA-TEST tests whether an LLM can replace one human annotation per item without reducing panel reliability by more than an acceptable margin. It uses a paired bootstrap: each replicate samples N items with replacement, draws k human annotations for each sampled item, replaces one of them with the LLM annotation, and compares the reliability of the human-only and mixed panels using a one-sided percentile-bootstrap non-inferiority test. Figure 1 illustrates the procedure and Appendix A provides the full algorithm.

Paired bootstrap. In each bootstrap iteration $b = 1, \ldots, B$, we sample N items with replacement from $\mathcal{D}$, denoted $\tilde{x}_1^b, \ldots, \tilde{x}_N^b$. For each sampled item $\tilde{x}_j^b$, let $\mathbf{y}_{j,h}^b$ be a panel of k human annotations drawn without replacement from its available annotations. Let $\mathbf{y}_{j,m}^b$ be formed by replacing one annotation in $\mathbf{y}_{j,h}^b$ with the LLM annotation for that item. Stacking the human-only and mixed panels gives $\mathbf{Y}_h^b$ and $\mathbf{Y}_m^b$, respectively. Then, given an adaptive margin ε^b defined below, we compute:

$$R_h^b = R(\mathbf{Y}_h^b), \qquad R_m^b = R(\mathbf{Y}_m^b), \qquad \widehat{\Delta}^b = R_m^b - R_h^b, \qquad \widehat{\Delta}_{\text{adj}}^b = \widehat{\Delta}^b + \varepsilon^b.$$

Adaptive margin. The ALTRA-TEST permits a small loss in the IAA score of the mixed panel. This reflects the practical trade-off behind replacement: LLM annotations are cheaper and easier to obtain than additional human annotations, so replacement can be useful even when it causes a small decrease in agreement. We express the largest acceptable loss as a non-inferiority margin ε^b. Rather than using the same fixed margin for all datasets, we set the margin adaptively. In well-defined or objective annotation tasks, human annotators are expected to agree strongly, and replacement should preserve this high reliability. In harder or more subjective tasks, human agreement is naturally lower, so the margin accommodates greater fluctuation in the mixed-panel agreement. We therefore scale the margin by the observed human disagreement, $1 - R_h^b$:

$$\varepsilon^b = c\frac{k_c}{k}\left(1 - R_h^b\right).$$

This margin has a direct interpretation. The constant c is the largest proportional increase in human disagreement that the LLM is allowed to introduce at the calibration panel size. In this paper, we calibrate c at $k_c = 3$ and set $c = 0.1$, so the mixed panel may increase disagreement by at most 10%. The factor k_c/k extends this margin to other panel sizes. This correction is needed because the effect of replacing one annotation scales as $O(1/k)$, i.e., diminishes as k grows: replacing one annotation out of three changes R_m more than replacing one annotation out of ten. We derive this scaling in Appendix B.2. Multiplying by k_c/k keeps the meaning of c comparable across panel sizes.

Statistical test. The ALTRA-TEST is a non-inferiority percentile-bootstrap test (Efron & Tibshirani, 1993, Eq. 13.5, p. 171). The null hypothesis is that replacing one human annotation with the LLM reduces agreement by more than the allowed margin:

$$H_0 : R_m + \varepsilon \leq R_h \quad \text{(i.e., } \Delta + \varepsilon \leq 0\text{)} \quad \textbf{vs.} \quad H_1 : R_m + \varepsilon > R_h \quad \text{(i.e., } \Delta + \varepsilon > 0\text{)}.$$

Let q^*_α be the empirical α-quantile of the adjusted bootstrap gaps $\{\widehat{\Delta}^b_{\text{adj}}\}_{b=1}^B$. Throughout the paper we use $\alpha = 0.05$. We reject H_0 and conclude that the LLM passes the ALTRA-TEST when $q^*_\alpha > 0$.

Comparing judges. While the test gives a binary replacement decision, the same bootstrap estimates naturally support comparison between LLM judges, regardless of whether they pass the test. We suggest comparing judges by the mean mixed-panel agreement. It should be used to compare judges within the same dataset, and read it alongside the human-only agreement R_h.

$$\overline{R}_m = \frac{1}{B}\sum_{b=1}^{B} R^b_m.$$

Consistency and validity. We theoretically establish that the ALTRA-TEST is a consistent test of asymptotic level α. Consistency means that q^*_α converges to Δ_{adj}, and that the probability of correctly passing a valid replacement asymptotically approaches 1. Asymptotic level α means that when the LLM is not a valid replacement, the probability of falsely passing it is asymptotically bounded by α. Under the standard regularity conditions stated in Appendix B, it follows that:

$$q^*_\alpha \xrightarrow{p} \Delta_{\text{adj}}, \qquad \Delta_{\text{adj}} \le 0 \Rightarrow \limsup_{N\to\infty} \Pr(q^*_\alpha > 0) \le \alpha, \quad \text{and} \quad \Pr(q^*_\alpha > 0) \to \begin{cases} 0, & \Delta_{\text{adj}} < 0, \\ 1, & \Delta_{\text{adj}} > 0, \end{cases}$$

Thus, as the number of items grows, the test rejects invalid replacements and passes valid replacements with probability approaching one, while controlling false passes under the null. The formal theorem and its proof are provided in Appendix B. The proof follows standard percentile-bootstrap arguments, adapted to account for two test-specific cases: the IAA-based statistic rather than a simple mean, and, within each bootstrap, within-item sampling of annotations without replacement.

4 SIMULATION-BASED ANALYSIS

Typical simulation studies of statistical tests examine finite-sample behavior by estimating Type I error and power under data drawn from the test's null and alternative hypotheses. For the ALTRA-TEST, the null and alternative are defined in terms of IAA scores. IAA scores are only a proxy for answering the replacement question: they measure whether the LLM preserves the annotation panel's reliability, not whether the LLM is closer than a human annotator to the ground-truth label. However, since ground-truth labels in real-world settings are unavailable (this is why human annotators are needed in the first place), annotation studies use IAA as a proxy for data quality.

In this section, we conduct simulations that go beyond this proxy and beyond Type I error and power, and answer the replacement question directly. We generate the ground-truth label of each item and control the quality of both the humans and the model. We therefore know whether the model is more or less likely than the humans to provide the ground-truth label, and whether it provides a better alternative. Instead of Type I error and power, we use measures that resemble them but better represent replacement decisions. We define the better-model pass rate and the worse-model pass rate: a good test should pass models that are less noisy than the humans (better-model) and reject models that are noisier (worse-model). These measures are defined by ground-truth accuracy rather than by the test's null and alternative hypotheses, which serve only as proxies. This also gives a common practical criterion for comparing procedures with different hypotheses.

4.1 EXPERIMENTAL SETUP

Data generation. We simulate annotation tasks with a known ground-truth label for each item. For each simulated dataset, we first generate the true item labels and then generate human and LLM annotations by adding noise. The human annotators share a noise level, η_h, and the LLM has its own noise level, η_f. Lower noise means a higher probability of providing the ground-truth label, so $\eta_f < \eta_h$ defines a better-model case and $\eta_f > \eta_h$ defines a worse-model case.

Following Calderon et al. (2025b), each simulation begins by drawing the prior probability of each label, and then drawing a ground-truth label for each item. This produces a synthetic dataset with known ground-truth labels. We then generate human and LLM annotations as follows: with probability $1-\eta$, the annotator returns the ground-truth label; with probability η, it returns a label sampled

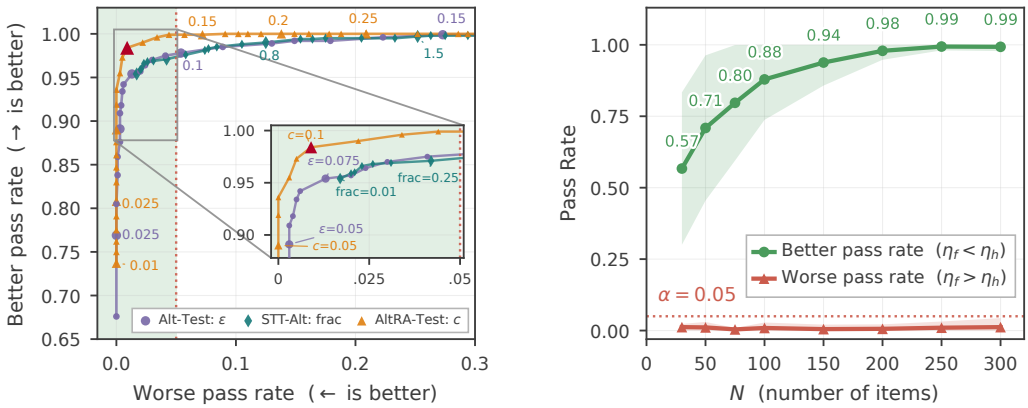


Figure 2: *Left:* The ALTRA-TEST against the baselines, simulated with $N = 200$ items and $k = 3$, varying each procedure's hyperparameter: c, ε, or frac. The X-axis shows the worse-model pass rate and the Y-axis the better-model pass rate. The green region marks the valid regime, where the worse-model pass rate is below 0.05. As shown, the ALTRA-TEST dominates the baselines. *Right:* Using the calibration in the left panel, we set $c = 0.1$ and vary the number of items N. From $N = 100$ on, the ALTRA-TEST achieves reliable better-model pass rates (> 0.85).

from the same label prior. In the main simulation, we draw the label prior from a Dirichlet distribution and generate a corrupted annotation as follows:

$$\pi \sim \text{Dirichlet}(\mathbf{1}_L), \qquad y_i \sim \text{Categorical}(\pi), \qquad \tilde{y}_{i,\eta} \sim \text{Categorical}\big((1-\eta)e_{y_i} + \eta\pi\big),$$

where L is the number of labels and e_{y_i} is the one-hot vector for the ground-truth label. Human annotations are generated as $h_{ij} = \tilde{y}_{i,\eta_h}$, while the LLM annotation is generated as $f_i = \tilde{y}_{i,\eta_f}$.

Parameters. In the main simulation, we use $L = 4$ labels, $m = 6$ human annotators per item, and $k = 3$ human annotations in the ALTRA-TEST. We vary η_h and η_f over $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, giving cases where the model is better than, worse than, or equal to the human annotators, with human agreement spanning 0.14 to 0.87. For each pair of noise levels, we generate multiple datasets and item subsamples, and measure agreement using Krippendorff's α. Unless otherwise stated, we use $N = 200$ items and test at $\alpha = 0.05$. Additional simulations with other generative distributions (uniform and Gaussian), group sizes (k), annotator-pool sizes (m), label-space sizes (L), and agreement metrics (Fleiss' κ, percent agreement, MSE, and RMSE) are reported in Appendix A.3.

Baselines. We compare the ALTRA-TEST with the two known statistical procedures closest to the replacement question. Both procedures require the same annotators to annotate all items. While our test can operate in a more relaxed setting, our simulated data satisfy their shared-annotator requirement and enable an apples-to-apples comparison. The first baseline is the ALT-TEST (Calderon et al., 2025b). It uses a leave-one-annotator-out procedure: each annotator is held out in turn, and a statistical test compares whether the LLM is non-inferior, up to a margin ε, to the held-out annotator in its alignment with the remaining annotators. The second procedure is STT (He et al., 2025a), which uses a bootstrap over items. It first computes the IAA of the all-human panel. Then, for each annotator, it replaces all their annotations with the LLM ones and computes the mean IAA of the resulting mixed panels. It uses Two One-Sided Tests (TOST; Rogers et al., 1993) to test whether the all-human and mixed-panel IAAs are equivalent, i.e., within a margin defined by a frac parameter. We evaluate STT-Alt, a non-inferiority adaptation of STT that uses only the lower one-sided test and for completeness, we also compare the ALTRA-TEST adjusted for equivalence testing with the original STT. Appendix A.2 provides full implementation details for all baselines and adaptations.

4.2 SIMULATION RESULTS

ALTRA-TEST gives the best replacement trade-off. We first calibrate the margin parameter for each test: c for ALTRA-TEST, ε for ALT-TEST, and frac for STT-ALT. Figure 2a compares the

procedures at $N = 200$ by sweeping these parameters and plotting the better-model pass rate against the worse-model pass rate. We focus on the valid region, where the worse-model pass rate is below $\alpha = 0.05$. In this region, the ALTRA-TEST gives a Pareto-dominant trade-off over the other tests: at the same worse-model pass rate, it passes more better-models than the baselines. This means that in the simulated settings, the ALTRA-TEST is more likely to pass models that truly improve over the human annotators. Appendix C shows the same pattern for sample sizes $N \in \{50, 100, 150, 300\}$ and under the uniform and Gaussian generative families. Finally, for the equivalence question, Figure 5 shows that the ALTRA-TEST adjusted for equivalence gives nearly the same trade-off as STT, and is slightly better in high worse-model regimes.

The allowance constant $c = 0.1$ is conservative and effective. As shown in Figure 2a, $c = 0.1$ lies near the knee of the ALTRA-TEST curve and remains inside the valid region; see also Figure 6a. At this value, the better-model pass rate is 0.98, while the worse-model pass rate remains below 0.05. The same value also performs well across other sample sizes and generative families. Since $c = 0.1$ has a direct interpretation, permitting a loss of agreement equal to 10% of the observed human disagreement, we recommend it as the default and use it throughout the real-world experiments.

More items mainly improve better-model detection. Figure 2b shows the behavior of the ALTRA-TEST as the number of items N increases, using $c = 0.1$. Like traditional power, adding items mainly improves the better-model pass rate: it reaches 0.88 at $N = 100$, 0.94 at $N = 150$, and 0.98 at $N = 200$. The worse-model pass rate remains below $\alpha = 0.05$ across the examined range. We therefore use $N = 200$ in the real-world experiments.

Analysis of other factors. Additional analyses at $c = 0.1$ are reported in the appendix. Across all examined annotator-pool sizes m, group sizes k, and label-space sizes L, the worse-model pass rate remains below 0.05 (Figures 7–9). Larger m stabilizes, and larger L increases, the better-model pass rate; larger k also provides modest gains, although $k = 3$ remains effective and is the most annotation-efficient choice. Among agreement metrics, Krippendorff's α and Fleiss' κ yield the highest better-model pass rates, whereas RMSE is the only metric for which the worse-model pass rate exceeds 0.05, an effect that can be mitigated by reducing c (Figure 10).

5 REAL-WORLD EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate the ALTRA-TEST on 20 datasets spanning text (12), image (5), audio (2), and video (1). The datasets cover nominal and ordinal tasks with two to ten labels, at least three annotations per item, and human agreement ranging from $\bar{R}_h = 0.02$ to 0.90. We use the a subset of 300 items per dataset and use only items with at least three human labels. Table 5 describes all 20 datasets.

We evaluate 18 LLMs as candidate annotators, assigning to each modality the models that natively accept its input: Qwen3-8B (Q3-8b) and Qwen3-14B (Q3-14b), Llama-3.2-3B (L3.2-3b) and Llama-3.1-8B (L3.1-8b), and Gemma-3-4B (G3-4b) and Gemma-3-12B (G3-12b) for text; the vision–language models Qwen2.5-VL-3B (Q2.5V-3b) and Qwen2.5-VL-7B (Q2.5V-7b), Llama-3.2-11B-Vision (L3.2V-11b) and the two Gemma-3 models for image, with the Qwen2.5-VL pair also handling video through sampled frames; and Qwen2.5-Omni-3B (Q2.5-O-3b) and Qwen2-Audio-7B (Q2-A-7b) for audio. Alongside these 11 open-weight models, spanning roughly 3B to 14B parameters, we evaluate seven proprietary judges: Gemini-3.1 Pro (Gem-P), Gemini-3.7 Flash (Gem-F) and Gemini-3.5 Flash-Lite (Gem-L) on all four modalities; GPT-5.6 Terra (GPT-T) and GPT-5.6 Luna (GPT-L) on text and image; and GPT-Audio-1.5 (GPT-A) and GPT-Audio-Mini (GPT-A-M) on audio. This gives 11 text, 10 image, 7 audio and 5 video judges. Dataset details , references and links for all models are provided in Appendix A.4.

Each judge is evaluated using up to five prompting techniques (Mizrahi et al., 2024; Li et al., 2025): zero-shot, few-shot, chain-of-thought (Wei et al., 2022), an ensemble of these three (Nahum et al., 2025), and GEPA prompt optimization (Agrawal et al., 2025). GEPA is applied only to open-weight models because of its computational cost. We run the ALTRA-TEST using panels of $k = 3$ annotations, Krippendorff's α as the agreement coefficient, $B = 500$ bootstrap iterations, a significance level of $\alpha = 0.05$, and the simulation-calibrated allowance constant $c = 0.1$.

Table 1: **Main results.** Rows are datasets with their human-only agreement $\bar{R}_h$; columns are LLM judges, grouped by family. Each cell reports the mean mixed-panel agreement $\bar{R}_m$. Cells in green mean the judge passes the ALTRA-TEST; **bold** marks the highest $\bar{R}_m$ for that dataset.

		Gemini			GPT		Qwen		Gemma		Llama	
Text	$\bar{R}_h$	**Gem-P**	**Gem-F**	**Gem-L**	**GPT-T**	**GPT-L**	**Q3-8b**	**Q3-14b**	**G3-4b**	**G3-12b**	**L3.2-3b**	**L3.1-8b**
DICES-350 (Aroyo et al., 2023)	0.20	0.14	0.06	0.10	0.16	0.16	0.09	0.16	0.12	**0.22**	0.14	0.15
EPIC (Frenda et al., 2023)	0.20	0.22	0.19	**0.24**	0.22	0.21	0.20	0.22	−0.06	0.19	0.01	0.02
MultiPref (Miranda et al., 2025)	0.28	**0.33**	0.30	0.28	0.24	0.24	0.19	0.26	0.14	0.25	0.20	0.12
ChaosNLI (Nie et al., 2020)	0.28	0.31	0.30	0.29	0.31	0.26	0.26	**0.32**	0.28	0.29	0.23	0.28
Persuasion	0.31	0.06	0.07	0.01	0.15	0.10	0.12	**0.18**	**0.18**	0.17	0.10	0.09
10k-prompts	0.37	0.40	0.36	**0.43**	0.35	0.35	0.39	0.15	0.33	0.39	0.23	0.28
Scruples (Lourie et al., 2021)	0.38	**0.45**	0.44	0.43	0.39	0.35	0.18	0.10	0.14	0.27	0.08	0.23
SummEval (Fabbri et al., 2021)	0.60	**0.65**	0.61	0.58	0.54	0.52	0.51	0.42	0.42	0.45	0.40	0.28
ConvAbuse (Curry et al., 2021)	0.72	**0.73**	0.71	**0.73**	0.68	0.71	0.70	0.65	0.41	0.61	0.48	0.58
CEBaB (Abraham et al., 2022)	0.72	**0.79**	**0.79**	**0.79**	**0.79**	**0.79**	0.78	0.77	0.78	**0.79**	0.78	0.78
Loneliness	0.90	0.57	**0.61**	0.43	0.38	0.44	0.18	0.39	0.22	0.38	0.33	0.22
Image & Video							**Q2.5V-3b**	**Q2.5V-7b**			**L3.2V-11b**	
T2I-DPO (Cai et al., 2025)	0.12	0.14	**0.15**	0.13	0.14	0.12	0.00	0.07	0.01	0.04	0.02	
Vibe-Arena	0.12	0.13	0.11	0.12	**0.14**	0.11	−0.05	0.05	−0.01	−0.00	−0.01	
FER+ (Barsoum et al., 2016)	0.51	**0.54**	0.52	0.46	0.49	0.49	0.28	0.31	0.36	0.38	0.42	
Flickr8k (Hodosh et al., 2013)	0.87	0.82	**0.85**	0.82	0.83	0.83	0.65	0.66	0.80	0.76	0.64	
CIFAR-10H (Peterson et al., 2019)	0.88	0.89	**0.90**	0.88	0.74	0.65	0.76	0.79	0.81	0.82	0.80	
T2V-Preference (He et al., 2025b)	0.04	**0.08**	**0.08**	0.06			−0.01	0.07				
Audio					**GPT-A**	**GPT-A-M**	**Q2.5-O-3b**	**Q2-A-7b**				
TTS-Preference	0.01	**0.02**	0.01	−0.00	−0.03	−0.00	0.01	−0.12				
SOMOS (Maniati et al., 2022)	0.05	−0.12	0.04	**0.05**	**0.05**	0.04	0.03	−0.00				

5.2 RESULTS

In almost half of the setups LLMs pass the test. Table 1 reports the default CoT setting for 18 LLMs and 20 datasets. It presents the post-substitution agreement scores, $\bar{R}_m$, and highlights in green the cases where the LLM passes the ALTRA-TEST. Under this setting, LLM judges pass in 71 of the 201 judge–dataset pairs (35.3%). When allowing each judge to use its best prompting technique, this number increases to 96 of 201 (47.8%; Appendix D). The outcome is strongly dataset-dependent, for example, all judges pass on CEBaB, whereas none pass on Persuasion, Measuring-Hate-Speech, Flickr8k-Expert, or Loneliness. Among the evaluated models, proprietary judges pass most often. Gemini-3.1-Pro passes 16/20 datasets (80%), while Gemini-3.7-Flash and Gemini-3.5-Flash-Lite pass 14/20 each (70%). GPT-5.6-Luna and GPT-5.6-Terra pass 9/17 (52.9%) and 8/17 (47.1%), respectively. The strongest open-weight judge, Gemma-3-12B, also passes 8/17 (47.1%).

Passing often means higher agreement. Under CoT, 49 of the 71 passing cells have $\bar{R}_m > \bar{R}_h$, and 59 have $\bar{R}_m \geq \bar{R}_h$. In these cases, replacing one human annotation with the LLM preserves or increases the agreement of the annotation panel, suggesting that the LLM is not merely an acceptable substitute but a better alternative to recruiting another annotator from the same pool. However, a higher $\bar{R}_m$ alone is not a replacement decision. In four CoT cases, $\bar{R}_m \geq \bar{R}_h$ but the judge still fails the test because the uncertainty is too large. Conversely, 12 passing cells have $\bar{R}_m < \bar{R}_h$, meaning that a small loss can still be statistically acceptable under the margin. The test is therefore needed to separate apparent changes in agreement from statistically supported replacement decisions.

Low human agreement limits what passing the test means. The datasets span a wide range of human agreement. Most of our generated-output preference datasets have $\bar{R}_h$ below 0.2, consistent with prior work showing that preferences are heterogeneous and weakly agreed upon (Calderon et al., 2025b;a; Nussbaum-Hoffer et al., 2026). Many subjective tasks, such as irony, safety, persuasiveness, and moral judgment, fall in a moderate range of $\bar{R}_h = 0.2$–0.6 (Lissak et al., 2024). Notably, $\bar{R}_h$ does not determine the result of the ALTRA-TEST: LLMs can pass or fail at both low and high agreement levels. The ALTRA-TEST is relative to the observed human pool. When $\bar{R}_h$ is around 0.2 or higher, a passing result indicates that the LLM is aligned with a non-trivial signal in the human pool. When $\bar{R}_h$ is near zero, a pass supports cost-equivalent substitution, and using an

Table 2: **Prompting techniques and minimum item count.** For each judge, the table reports the number of datasets on which it passes under *at least one* prompting technique, the pass rate (in %) of each technique, and the average minimum number of items needed to pass among datasets where it passes at $N = 300$. **Bold** marks the judge's best prompting technique.

	Gemini			GPT		Qwen				Gemma		Llama	
	Gem-P	**Gem-F**	**Gem-L**	**GPT-T**	**GPT-L**	**Q3-8b**	**Q3-14b**	**Q2.5V-3b**	**Q2.5V-7b**	**G3-4b**	**G3-12b**	**L3.2-3b**	**L3.1-8b**
# passed	16/20	14/20	14/20	8/17	9/17	4/12	5/12	1/6	3/6	2/17	8/17	2/12	2/12
Zero-shot	65%	55%	50%	**41%**	41%	33%	**42%**	**17%**	**33%**	12%	29%	8%	17%
Few-shot	**80%**	50%	50%	**41%**	29%	33%	33%	0%	**33%**	12%	24%	8%	17%
CoT	60%	**65%**	**55%**	35%	29%	33%	33%	0%	17%	12%	29%	8%	17%
Ensemble	70%	**65%**	50%	**41%**	**53%**	33%	**42%**	**17%**	17%	12%	**41%**	**17%**	17%
GEPA	✗	✗	✗	✗	✗	33%	33%	0%	0%	12%	29%	8%	17%
min N	100	111	96	122	133	75	70	225	117	75	147	138	88

LLM will not change much. This is the key nuance: the ALTRA-TEST evaluates replacement, not whether the annotations contain a useful signal. If the human pool does not exhibit such a signal, replacement, although justified, may not be meaningful. In such cases, we recommend first obtaining annotations with at least fair agreement ($\bar{R}_h > 0.2$) before testing for replacement.

Prompting helps, but rarely changes the decision. Prompting changes some verdicts, and its effect depends on both the model and the dataset. Overall, ensemble prompting, i.e., majority vote over different prompts, gives the highest aggregate pass count: 84 of 201 (41.8%), compared with 38.3% for zero-shot, 36.8% for few-shot, and 35.3% for CoT. GEPA, evaluated only for open-weight models, does not increase replacement decisions (19.1% for GEPA compared with 20.2% for CoT when considering the same models). Furthermore, in Appendix D, we report additional experiments varying the number of few-shot examples and the CoT reasoning effort. We find that more examples or more reasoning can help, but the trends are mixed and dataset-dependent.

About 100 items often suffice to pass. We next ask how many items are required for reliable ALTRA-TEST results in real-world settings. Unlike simulations, real data do not provide ground truth about whether an LLM should pass. We therefore use the 300-item experiments as the reference verdict and ask, for LLMs that pass with 300 items, what was the minimum number of items for which they pass. Table 2 reports these minimum passing sizes. Across judges, the median is 111 items and the mean is 115, with all values between 70 and 150 except for Qwen2.5-VL-3B, which reaches 225. The full sample-size analysis is provided in Appendix D, showing that decisions often stabilize around 100 items and that this is enough for strong judges to pass.

6 CONCLUSION

As LLMs are increasingly used in place of human annotators, their efficiency must be accompanied by statistical evidence that the replacement preserves annotation quality. We propose ALTRA-TEST, the Alternative Annotator Test with Resampled Annotations, which formulates this as a non-inferiority test: it asks whether the inter-annotator agreement (IAA) of a mixed human-LLM panel is statistically non-inferior to that of an all-human panel. Unlike prior replacement tests, ALTRA-TEST can be applied to sparse annotation settings, where different items are labeled by different subsets of annotators and no shared panel is required.

Across simulations and real-world experiments, ALTRA-TEST achieves a stronger trade-off than existing procedures: for the same rate of false replacements, it passes more valid ones. Around 100 annotated items often suffice for stable decisions. Empirically, replacement is possible in many settings, but the outcome depends more on the dataset than on the judge or prompting technique. The main implication is that LLM annotators should be validated per task rather than adopted as general practice. Because the test is relative to the observed human pool, its verdict should always be reported alongside human IAA: when human agreement is itself very low, a passing result does not show that the labels contain a useful signal. We hope ALTRA-TEST helps researchers make more rigorous and transparent decisions about when LLMs can replace human annotators.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, provide critical ingredients for proving mathematical claims, and assisting in writing proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, or interpreting results, and generating synthetic data sets, assisting with translation, and cleaning and reformatting datasets are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, drafting parts of a research paper, summarizing or analyzing existing literature, and editing the research paper to improve readability. We have reviewed all AI-assisted work. We checked LLM-generated theoretical proofs with 3 authors, and LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the implementation of the ALTRA-TEST, experimental data, and supporting materials in the anonymous repository at 🤗 `anonymous-hf.com/a/6y60hp9vqg6k`. Sections 3 and 5 formally define the procedure and report all test parameters, while Algorithm 1 specifies the complete bootstrap process. Section 4 and Appendix A.3 describe the simulation protocol, calibration of the allowance constant, comparisons with existing tests, and robustness analyses. The evaluated datasets, their preprocessing, and annotation statistics are documented in Appendix A.4; full model and prompting configurations are reported in Section 5.1 and Appendix A.4; and complete results for every dataset, judge, and prompting technique are provided in Appendix D. Together, these materials contain the information required to reproduce the simulations, empirical experiments, figures, and tables reported in this work.

REFERENCES

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on NLP model behavior. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/701ec28790b29a5bc33832b7bdc4c3b6-Abstract-Conference.html`.

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Daniel Klein, Matei Zaharia, and Omar Khattab. GEPA: reflective prompt evolution can outperform reinforcement learning. *CoRR*, abs/2507.19457, 2025. doi: 10.48550/ARXIV.2507.19457. URL `https://doi.org/10.48550/arXiv.2507.19457`.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *CoRR*, abs/2301.09633, 2023. doi: 10.48550/ARXIV.2301.09633. URL `https://doi.org/10.48550/arXiv.2301.09633`.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36(1):15–24, 2015. doi: 10.1609/AIMAG.V36I1.2564. URL `https://doi.org/10.1609/aimag.v36i1.2564`.

Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. DICES dataset: Diversity in conversational AI evaluation for safety. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/`

2023/hash/a74b697bce4cac6c91896372abaa8863-Abstract-Datasets_and_Benchmarks.html.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguistics*, 34(4):555–596, 2008. doi: 10.1162/COLI.07-034-R2. URL https://doi.org/10.1162/coli.07-034-R2.

Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Yukiko I. Nakano, Elisabeth André, Toyoaki Nishida, Louis-Philippe Morency, Carlos Busso, and Catherine Pelachaud (eds.), *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, pp. 279–283. ACM, 2016. doi: 10.1145/2993148.2993165. URL https://doi.org/10.1145/2993148.2993165.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pp. 15–21, 2021.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. Llms instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 238–255. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-SHORT.20. URL https://doi.org/10.18653/v1/2025.acl-short.20.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas J. Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18434–18443. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01718. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cai_Diffusion_Self-Distillation_for_Zero-Shot_Customized_Image_Generation_CVPR_2025_paper.html.

Nitay Calderon, Liat Ein-Dor, and Roi Reichart. Multi-domain explainability of preferences. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 14542–14575. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.EMNLP-MAIN.736. URL https://doi.org/10.18653/v1/2025.emnlp-main.736.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 16051–16081. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.ACL-LONG.782. URL https://doi.org/10.18653/v1/2025.acl-long.782.

David Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 15607–15631. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.870. URL https://doi.org/10.18653/v1/2023.acl-long.870.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report.

CoRR, abs/2407.10759, 2024. doi: 10.48550/ARXIV.2407.10759. URL https://doi.org/10.48550/arXiv.2407.10759.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. Convabuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. *CoRR*, abs/2109.09483, 2021. URL https://arxiv.org/abs/2109.09483.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Trans. Assoc. Comput. Linguistics*, 10:92–110, 2022. doi: 10.1162/TACL_A_00449. URL https://doi.org/10.1162/tacl_a_00449.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is GPT-3 a good data annotator? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 11173–11195. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.626. URL https://doi.org/10.18653/v1/2023.acl-long.626.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Springer, 1993. ISBN 978-1-4899-4541-9. doi: 10.1007/978-1-4899-4541-9. URL https://doi.org/10.1007/978-1-4899-4541-9.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d862f7f5445255090de13b825b880d59-Abstract-Conference.html.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=E8gYIrbP00.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409, 2021. doi: 10.1162/TACL_A_00373. URL https://doi.org/10.1162/tacl_a_00373.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of llms with imperfect judges. *CoRR*, abs/2601.20913, 2026. doi: 10.48550/ARXIV.2601.20913. URL https://doi.org/10.48550/arXiv.2601.20913.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6715–6726. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.415. URL https://doi.org/10.18653/v1/2023.emnlp-main.415.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. EPIC: multi-perspective annotation of a corpus of irony. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13844–13857. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.774. URL https://doi.org/10.18653/v1/2023.acl-long.774.

Team Gemma. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025. doi: 10.48550/ARXIV.2503.19786. URL https://doi.org/10.48550/arXiv.2503.19786.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056, 2023. doi: 10.48550/ARXIV.2303.15056. URL https://doi.org/10.48550/arXiv.2303.15056.

Kristina Gligoric, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident LLM annotations be used for confident conclusions? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 3514–3533. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.179. URL https://doi.org/10.18653/v1/2025.naacl-long.179.

Google DeepMind. Gemini. https://deepmind.google/models/gemini, 2026.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642, 2025.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna M. Wallach, Steven Z. Wu, and Alexandra Chouldechova. Validating llm-as-a-judge systems under rating indeterminacy. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a309239c11a28c597d050bd4a1752d32-Abstract-Conference.html.

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.

Jiaman He, Zikang Leng, Dana McKay, Damiano Spina, and Johanne R. Trippas. Can we hide machines in the crowd? quantifying equivalence in llm-in-the-loop annotation tasks. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2025, Xi'an, China, December 7-10, 2025*, pp. 426–436. ACM, 2025a. doi: 10.1145/3767695.3769508. URL https://doi.org/10.1145/3767695.3769508.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. Videoscore2: Think before you score in generative video evaluation. *CoRR*, abs/2509.22799, 2025b. doi: 10.48550/ARXIV.2509.22799. URL https://doi.org/10.48550/arXiv.2509.22799.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.*, 47:853–899, 2013. doi: 10.1613/JAIR.3994. URL https://doi.org/10.1613/jair.3994.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. Analyzing dataset annotation quality management in the wild. *Comput. Linguistics*, 50(3):817–866, 2024. doi: 10.1162/COLI_A_00516. URL https://doi.org/10.1162/coli_a_00516.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report llm-as-a-judge evaluations. *CoRR*, abs/2511.21140, 2025. doi: 10.48550/ARXIV.2511.21140. URL https://doi.org/10.48550/arXiv.2511.21140.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 2757–2791. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.138. URL https://doi.org/10.18653/v1/2025.emnlp-main.138.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 2040–2079. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.113. URL https://doi.org/10.18653/v1/2024.naacl-long.113.

Team Llama. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. SCRUPLES: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 13470–13479. AAAI Press, 2021. doi: 10.1609/AAAI.V35I15.17589. URL https://doi.org/10.1609/aaai.v35i15.17589.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. SOMOS: the samsung open MOS dataset for the evaluation of neural text-to-speech synthesis. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 2388–2392. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-10922. URL https://doi.org/10.21437/Interspeech.2022-10922.

Katelyn Xiaoying Mei, Yi-Li Hsu, Minjoon Choi, Zongwan Cao, Chenjun Xu, Bingbing Wen, Su Lin Blodgett, and Lucy Lu Wang. Illusions of the gold standard: A large-scale analysis of human evaluation protocols for long-form text generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 13939–13961. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.635. URL https://doi.org/10.18653/v1/2026.acl-long.635.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. AI feedback. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 7162–7200. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.355. URL https://doi.org/10.18653/v1/2025.acl-long.355.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Trans. Assoc. Comput. Linguistics*, 12:933–949, 2024. doi: 10.1162/TACL_A_00681. URL https://doi.org/10.1162/tacl_a_00681.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. Are llms better than reported? detecting label errors and mitigating their effect on model performance.

In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 26782–26809. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1360. URL https://doi.org/10.18653/v1/2025.emnlp-main.1360.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 9131–9143. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.734. URL https://doi.org/10.18653/v1/2020.emnlp-main.734.

Nirit Nussbaum-Hoffer, Nitay Calderon, Liat Ein-Dor, and Roi Reichart. LLM explainability with counterfactual chains and causal graphs. *CoRR*, abs/2606.05972, 2026. doi: 10.48550/ARXIV.2606.05972. URL https://doi.org/10.48550/arXiv.2606.05972.

Nagham Omar, Maya Rozenshtein, Evgeny Mishlyakov, and Avigdor Gal. Train smarter, not harder: Switching signal-guided training in active learning. *arXiv preprint arXiv:2609.06806*, 2026.

OpenAI. GPT-5.6. https://openai.com/index/gpt-5-6, 2026a.

OpenAI. GPT-Audio-1.5. https://developers.openai.com/api/docs/models/gpt-audio-1.5, 2026b.

OpenAI. GPT-Audio-Mini. https://developers.openai.com/api/docs/models/gpt-audio-mini, 2026c.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing bayesian models of annotation. *Trans. Assoc. Comput. Linguistics*, 6:571–585, 2018. doi: 10.1162/TACL_A_00040. URL https://doi.org/10.1162/tacl_a_00040.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694, 2019. doi: 10.1162/TACL_A_00293. URL https://doi.org/10.1162/tacl_a_00293.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9616–9625. IEEE, 2019. doi: 10.1109/ICCV.2019.00971. URL https://doi.org/10.1109/ICCV.2019.00971.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 10671–10682. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.731. URL https://doi.org/10.18653/v1/2022.emnlp-main.731.

Team Qwen. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025a. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Team Qwen. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025b. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, 2010. doi: 10.5555/1756006.1859894. URL https://dl.acm.org/doi/10.5555/1756006.1859894.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

James L Rogers, Kenneth I Howard, and John T Vessey. Using significance tests to evaluate equivalence between two experimental groups. *Psychological bulletin*, 113(3):553, 1993.

Pratik S. Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris J. Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma (eds.), *Proceedings of the 1st Workshop on Perspectivist Approaches to NLPerspectives@LREC 2022, Marseille, France, 20th June 2022*, pp. 83–94. European Language Resources Association, 2022. URL https://aclanthology.org/2022.nlperspectives-1.11.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 5884–5906. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.431. URL https://doi.org/10.18653/v1/2022.naacl-main.431.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the llm-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In Toby Li, Fabio Paternò, Kaisa Väänänen, Luis Leiva, Lucio Davide Spano, and Katrien Verbert (eds.), *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI 2025, Cagliari, Italy, March 24-27, 2025*, pp. 952–966. ACM, 2025. doi: 10.1145/3708359.3712091. URL https://doi.org/10.1145/3708359.3712091.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 930–957. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.54. URL https://doi.org/10.18653/v1/2024.emnlp-main.54.

Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *CoRR*, abs/2304.06588, 2023. doi: 10.48550/ARXIV.2304.06588. URL https://doi.org/10.48550/arXiv.2304.06588.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. Evaluating large language models as expert annotators. *CoRR*, abs/2508.07827, 2025. doi: 10.48550/ARXIV.2508.07827. URL https://doi.org/10.48550/arXiv.2508.07827.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470, 2021. doi: 10.1613/JAIR.1.12752. URL https://doi.org/10.1613/jair.1.12752.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat. Mac. Intell.*, 7(3):400–411, 2025. doi: 10.1038/S42256-025-00986-Z. URL https://doi.org/10.1038/s42256-025-00986-z.

Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of llm-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. *CoRR*, abs/2408.13006, 2024. doi: 10.48550/ARXIV.2408.13006. URL https://doi.org/10.48550/arXiv.2408.13006.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 297–312. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.21/.

Xiaobao Wu. A comprehensive survey on learning from rewards for large language models: Reward models and learning strategies. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 17847–17875. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.970. URL https://doi.org/10.18653/v1/2025.findings-emnlp.970.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025. doi: 10.48550/ARXIV.2503.20215. URL https://doi.org/10.48550/arXiv.2503.20215.

Michael J. Q. Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhang25bx.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3623–3643. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.198. URL https://doi.org/10.18653/v1/2024.acl-long.198.

Caleb Ziems, William Barr Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Comput. Linguistics*, 50(1):237–291, 2024. doi: 10.1162/COLI_A_00502. URL https://doi.org/10.1162/coli_a_00502.

APPENDIX CONTENTS

A Additional Implementation Details **18**

A.1 EQ-ALTRA-TEST: Testing for Equivalence 19

A.2 ALT-TEST and STT 19

A.3 Simulations 21

A.4 Datasets, Models, and Prompting 21

B Theoretical Guarantees **22**

B.1 Population Quantities 22

B.2 Panel-Size Scaling 23

B.3 Consistency and Asymptotic level α 25

C Simulations: Additional Results **26**

D Real-World Data: Additional Results **31**

A ADDITIONAL IMPLEMENTATION DETAILS

Table 3: Main notation used in the paper.

Notation	Meaning
$\mathcal{D} = \{x_1, \ldots, x_N\}$	Dataset of N items.
$H_i = \{h_{i,1}, \ldots, h_{i,M_i}\}$	The M_i human annotations observed for item x_i.
$f(x_i)$	LLM annotation for item x_i.
$k,\ k_c$	Panel size used by the test, with $k \geq 3$ and $M_i \geq k$, and the panel size at which c is calibrated.
c	Allowance constant; $c = 0.1$ by default.
$\mathbf{y}_{i,h},\ \mathbf{y}_{i,m}$	Human-only panel of k annotations for item x_i, and the mixed panel formed by replacing one of them with $f(x_i)$; stacking them over items gives the matrices $\mathbf{Y}_h$ and $\mathbf{Y}_m$.
$R(\cdot)$	Chance-corrected agreement score; higher values indicate more reliable panels.
$R_h = R(\mathbf{Y}_h),\ R_m = R(\mathbf{Y}_m)$	Human-only and mixed-panel agreement.
$\Delta = R_m - R_h$	Substitution gap.
$\varepsilon = c\frac{k_c}{k}(1 - R_h)$	Adaptive non-inferiority margin.
$\Delta_{\text{adj}} = \Delta + \varepsilon$	Adjusted substitution gap; the null hypothesis is $\Delta_{\text{adj}} \leq 0$.
B	Number of bootstrap iterations.
$R_h^{(b)},\ R_m^{(b)}$	Human-only and mixed agreement in bootstrap iteration b; $\varepsilon^{(b)}$, $\widehat{\Delta}^{(b)}$ and $\widehat{\Delta}_{\text{adj}}^{(b)}$ are defined analogously.
q_α^*	Empirical α-quantile of the adjusted bootstrap gaps; the LLM passes when $q_\alpha^* > 0$.
$\bar{R}_h,\ \bar{R}_m$	Mean human-only and mixed-panel agreement across bootstrap iterations; $\bar{R}_m$ is used to compare judges.
L	Number of labels.
$m,\ \eta_h,\ \eta_f$	In the simulations: human annotators per item, and the noise levels of the humans and of the LLM.

Algorithm 1 ALTRA-TEST paired bootstrap

Input: dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$, human annotations H_i, LLM annotations $f(x_i)$, panel size k, calibrated panel size k_c, allowance constant c, agreement score R, bootstrap iterations B, significance level α.
for $b = 1, \ldots, B$ **do**
  Sample N items with replacement from $\mathcal{D}$, denoted $\tilde{x}_1^{(b)}, \ldots, \tilde{x}_N^{(b)}$.
  for each sampled item $\tilde{x}_j^{(b)}$ **do**
    Draw without replacement k human annotations of $\tilde{x}_j^{(b)}$, forming $\mathbf{y}_{j,h}^{(b)}$.
    Form the mixed panel $\mathbf{y}_{j,m}^{(b)}$ by replacing one random annotation with $f(\tilde{x}_j^{(b)})$.
  end for
  Stack the panels into $\mathbf{Y}_h^{(b)}$ and $\mathbf{Y}_m^{(b)}$, and compute:

$$R_h^{(b)} = R(\mathbf{Y}_h^{(b)}), \qquad R_m^{(b)} = R(\mathbf{Y}_m^{(b)}).$$

  Compute the adjusted substitution gap:

$$\widehat{\Delta}_{\text{adj}}^{(b)} = \widehat{\Delta}^{(b)} + \varepsilon^{(b)} = \big(R_m^{(b)} - R_h^{(b)}\big) + c\frac{k_c}{k}\big(1 - R_h^{(b)}\big).$$

end for
Let q_α^* be the empirical α-quantile of $\{\widehat{\Delta}_{\text{adj}}^{(b)}\}_{b=1}^B$.
Compute $\bar{R}_m = B^{-1}\sum_{b=1}^B R_m^{(b)}$.
Return: binary decision $\mathbf{1}\{q_\alpha^* > 0\}$ and comparison score $\bar{R}_m$.

A.1 EQ-ALTRA-TEST: TESTING FOR EQUIVALENCE

The ALTRA-TEST tests non-inferiority: it asks whether the substitution costs less agreement than the allowance permits, so a model that improves the panel passes comfortably. Some settings call for the stricter question of whether the model is *indistinguishable* from a human annotator, so that a judge raising agreement fails alongside one lowering it. This arises where the annotations are meant to reflect human judgment rather than correctness, and labels more consistent than the humans' would misrepresent the population being measured.

The procedure needs one change. Where Section 3.1 rejects H_0 when the α-quantile of the adjusted gaps $\widehat{\Delta}^{(b)} + \varepsilon^{(b)}$ clears zero, the equivalence version requires the whole interval to lie inside the allowance:

$$H_0 : |\Delta| \geq \varepsilon \qquad \text{versus} \qquad H_1 : |\Delta| < \varepsilon, \tag{1}$$

tested by

$$q_\alpha^* > 0 \quad \wedge \quad \tilde{q}_{1-\alpha}^* < 0, \tag{2}$$

where $\tilde{q}_{1-\alpha}^*$ is the $(1-\alpha)$-quantile of $\widehat{\Delta}^{(b)} - \varepsilon^{(b)}$ over the same bootstrap replicates. This is the two one-sided tests procedure written as interval inclusion, with the allowance formed inside each replicate as in the one-sided rule, and it reuses everything else unchanged, so both versions can be run from the same bootstrap replicates. We refer to this variant as the EQ-ALTRA-TEST.

Two things change with it. The constant c must be recalibrated, since $c = 0.1$ was chosen for the one-sided rule and adding the upper bound reduces the better-model pass rate from 0.984 to 0.030 (Appendix C); choose it against the tolerance the setting calls for, as with $c = 0.125$ in Appendix C. And the verdict should be read differently, since a failure no longer says the model is worse, only that it is distinguishable, which may be because it is better.

A.2 ALT-TEST AND STT

Two existing procedures address closely related replacement decisions. The ALT-TEST (Calderon et al., 2025b) compares the LLM separately with each human annotator, and the Statistical Turing Test (STT; He et al., 2025a) evaluates whether LLM and human annotations are statistically indistinguishable. Table 4 compares them with the ALTRA-TEST, and this subsection describes how we

Table 4: Comparison of the three procedures evaluated in the paper. ✓ marks a property the test has; ✗ marks a limitation or requirement.

	ALT-TEST (Calderon et al., 2025b)	**STT** (He et al., 2025a)	**ALTRA-TEST (ours)**
Question	can the LLM replace a human annotator?	is the LLM indistinguishable from a human annotator?	can the LLM replace a human annotator?
Hypothesis	non-inferiority, one-sided	equivalence, two-sided	non-inferiority, one-sided
Compared against	the excluded annotator, head to head	the all-human mean	the annotator pool
Annotator identities	✗ required	✗ required	✓ not required
Annotator structure	✗ repeated annotators across items	✗ two disjoint groups, same items	✓ a single pool, no overlap needed
Annotators per item	≥ 3	≥ 4 (≥ 2 per group)	≥ 3
Recommended items	50–100	≥ 80	≥ 100
Agreement metric	task-specific: ACC, $-$RMSE, SIM	Krippendorff's α	✓ any chance-corrected IAA
Test	paired t-test per annotator	TOST (two one-sided t-tests)	bootstrap quantile of $\Delta + \varepsilon$
Decision rule	$\omega \geq 0.5$	both one-sided $p < \alpha$	$q^*_\alpha > 0$
Margin	✗ fixed $\varepsilon \in \{0.1, 0.15, 0.2\}$, set by annotator type	✓ $(\alpha_a - \alpha_b) \cdot \mathrm{frac}$, factor unspecified	✓ $c\,(1 - R_h^{(b)})\,k_c/k$, per replicate, calibrated
Ranks several LLMs	✓ ρ (adv. probability)	✗	✓ $\bar{R}_m$

implement them. All procedures receive identical inputs whenever their data requirements permit: in each simulation condition, the dataset and item subsample are generated once and supplied to all procedures, and in the empirical analysis the tests receive, for each dataset, the items that satisfy the requirements of every compared test, together with the same LLM annotations. Differences in their verdicts can therefore be attributed to the procedures rather than to variation in the evaluated data.

ALT-TEST. We use the implementation provided by Calderon et al. (2025b). Each human annotator is excluded in turn, and the alignment of the LLM with the remaining annotators is compared with that of the excluded annotator. The annotator-level tests are corrected for multiple comparisons, and the LLM passes when it outperforms at least half of the annotator pool. Because the margin is determined by annotator type rather than by a single value, we evaluate a range of margins and mark the three recommended settings, $\varepsilon = 0.10$, 0.15, and 0.20 for crowd workers, skilled annotators, and experts. On real datasets the procedure applies only when annotator identities are available and each annotator labels sufficiently many items, so the empirical comparison is restricted to the datasets satisfying these conditions.

Statistical Turing Test (STT). We implement the STT following He et al. (2025a). Each member of an all-human panel is replaced by the LLM in turn, the resulting mixed-panel agreements are compared with the all-human agreement using a paired bootstrap, and equivalence is established when both one-sided tests reject. The equivalence margin is estimated from two disjoint human panels that annotate the same items. Because the original work illustrates several scaling factors without recommending one, we evaluate a range of factors and use $\mathrm{frac} = 0.5$ on real data. The two-panel requirement is satisfied in our simulations, but among our real datasets only DICES-350 satisfies it; we therefore evaluate the STT systematically in simulation and on real data only for DICES-350.

STT adapted to the alternative annotator question (STT-ALT). The STT evaluates equivalence, whereas the ALT-TEST and ALTRA-TEST evaluate non-inferiority. This complicates direct comparisons: conditions in which the LLM is more accurate than the humans are better-model conditions that a non-inferiority test should pass, but an equivalence test may correctly fail them because the LLM differs from the humans in the favorable direction. We therefore introduce STT-ALT, which adapts the STT to the alternative annotator question, that is, to non-inferiority. STT-ALT retains the original sampling, agreement estimation, bootstrap procedure, and margin, but applies only the lower one-sided test, asking whether the mixed panel is no worse than the all-human panel within the margin. No quantities are re-estimated: both verdicts are derived from the same bootstrap statistics, so every run that passes the STT also passes STT-ALT, whereas STT-ALT may additionally pass runs that violate only the upper equivalence bound.

A.3 SIMULATIONS

Replications. Each simulated dataset contains 600 items, generated as described in Section 4. The noise grid contains 25 conditions: 10 worse-model conditions, in which the LLM is noisier than the humans, 10 better-model conditions, in which it is less noisy, and 5 in which their noise levels are equal. For each condition, we generate 10 independent datasets and draw 10 independent subsamples of N items from each, without replacement. Subsamples drawn from the same dataset may overlap, but no item is repeated within a subsample. This design yields 100 test decisions per condition and 2500 decisions overall. Every run uses $B = 500$ bootstrap iterations, and because $k = k_c$ in the default setting, the allowance simplifies to $\varepsilon^{(b)} = c(1 - R_h^{(b)})$; the sweep over k applies the full k_c/k adjustment.

Generative families. The baseline simulation draws the class prior from a Dirichlet distribution and samples noisy labels from that prior, giving a skewed nominal distribution. We repeat the complete noise grid under two additional families that vary both the gold-label and the error distribution:

- **Uniform:** gold labels are uniformly distributed over the L categories, and noisy labels are sampled uniformly from the remaining $L - 1$ categories, giving a balanced nominal distribution.
- **Gaussian:** gold labels follow a discretized normal distribution, and a noisy label is generated as $\mathrm{round}(y + \mathcal{N}(0, \sigma))$, with values reflected at the boundaries and σ chosen to obtain the target marginal error rate η. This represents ordinal tasks in which errors occur between nearby categories.

A.4 DATASETS, MODELS, AND PROMPTING

Datasets. Table 5 lists all 20 datasets with their task, label type, number of labels L, usable items N, those carrying at least 3 human labels, the number of human annotations per item M_i, and human-only agreement $\bar{R}_h$. The collection mixes the two settings a practitioner encounters: some datasets carry persistent annotator identifiers, so that a worker can be followed across items, while others record only anonymous labels per item. The ALTRA-TEST treats both the same way, using only the M_i labels of each item.

Table 5: **Dataset descriptions.** The table reports the task, label type, number of labels, number of usable items, number of human annotations per item, and human-only agreement $\bar{R}_h$. Here, ord. means ordinal, nom. means nominal, L is the number of labels, N is the number of items with at least three labels, and M_i is the number of human annotations for item x_i.

Dataset	Task	Type	L	N	M_i	$\bar{R}_h$
Text						
EPIC (Frenda et al., 2023)	Irony detection	ord.	2	1500	5–8	0.20
DICES-350 (Aroyo et al., 2023)	Conversational safety	ord.	3	350	123	0.20
ChaosNLI (Nie et al., 2020)	NLI entailment	nom.	3	1500	100	0.28
Measuring-Hate-Speech (Sachdeva et al., 2022)	Hate-speech severity	ord.	3	1500	4–815	0.53

Table 5, continued

Dataset	Task	Type	L	N	M_i	$\bar{R}_h$
Loneliness[2]	Loneliness intensity	ord.	5	497	3	0.90
CEBaB (Abraham et al., 2022)	Restaurant star rating	ord.	5	711	3–4	0.72
MultiPref (Miranda et al., 2025)	Preference strength	ord.	5	1500	4	0.28
Scruples (Lourie et al., 2021)	Moral judgment	nom.	5	1500	30	0.38
SummEval (Fabbri et al., 2021)	Summary quality	ord.	5	1600	3	0.60
10k-prompts[3]	Prompt quality rank	ord.	5	391	3–5	0.37
ConvAbuse (Curry et al., 2021)	Abuse severity	ord.	5	2225	3–8	0.72
Persuasion[4]	Argument persuasiveness	ord.	7	1313	3	0.31
Image						
T2I-DPO (Cai et al., 2025)	Text-to-image preference	nom.	2	400	8	0.12
Vibe-Arena[5]	Landing-page preference	nom.	2	400	15	0.12
Flickr8k-Expert (Hodosh et al., 2013)	Caption quality	ord.	4	400	3	0.87
FER+ (Barsoum et al., 2016)	Facial expression	nom.	8	600	10	0.51
CIFAR-10H (Peterson et al., 2019)	Image classification	nom.	10	600	30	0.88
Audio						
TTS-Preference[6]	TTS naturalness pref.	nom.	2	1000	15	0.01
SOMOS (Maniati et al., 2022)	Speech naturalness (MOS)	ord.	5	20100	16–27	0.05
Video						
T2V-Preference (He et al., 2025b)	Text-to-video preference	nom.	2	300	74–90	0.04

Models. The open-weight judges are Qwen3-8B and Qwen3-14B (Qwen, 2025b), Llama-3.2-3B, Llama-3.1-8B and Llama-3.2-11B-Vision (Llama, 2024), Gemma-3-4B and Gemma-3-12B (Gemma, 2025), Qwen2.5-VL-3B and Qwen2.5-VL-7B (Qwen, 2025a), Qwen2.5-Omni-3B (Xu et al., 2025), and Qwen2-Audio-7B (Chu et al., 2024). The proprietary judges are Gemini-3.1 Pro, Gemini-3.7 Flash and Gemini-3.5 Flash-Lite (Google DeepMind, 2026), GPT-5.6 Terra and GPT-5.6 Luna (OpenAI, 2026a), GPT-Audio-1.5 (OpenAI, 2026b) and GPT-Audio-Mini (OpenAI, 2026c). On video, the Qwen2.5-VL pair receives sampled frames, so the verdict applies to that complete processing pipeline rather than to the model alone.

Prompting. Each judge receives the task definition and the annotation guidelines of the dataset. Few-shot prompting uses five randomly sampled in-context examples, chain-of-thought uses mild reasoning effort, and the ensemble takes the majority vote over the zero-shot, few-shot and chain-of-thought labels. Appendix D reports how the verdicts change with the number of few-shot examples and with the reasoning effort.

B Theoretical Guarantees

This section establishes two properties of the AltRA-Test. First, the substitution gap decreases at rate $1/k$ as the panel size grows, supporting the panel-size adjustment in the adaptive allowance. Second, under standard regularity conditions, the lower bootstrap quantile consistently estimates the adjusted substitution gap and yields an consistent test of asymptotic level α.

B.1 Population Quantities

Let $Z_i = (X_i, H_i, F_i)$, $i = 1, \ldots, N$, denote the observed item clusters, where H_i is the multiset of available human labels and $F_i = f(X_i)$ is the label produced by a fixed model and prompt. We assume that every item has at least $k \geq 3$ human labels.

[2] https://huggingface.co/datasets/yael-katsman/Loneliness-Causes-and-Intensity
[3] https://huggingface.co/datasets/data-is-better-together/10k_prompts_ranked
[4] https://huggingface.co/datasets/Anthropic/persuasion
[5] https://huggingface.co/datasets/datapointai/vibe-landing-page-arena
[6] https://huggingface.co/datasets/datapointai/tts-human-preferences-medium

For each item, a panel draw U samples k distinct human annotations uniformly without replacement and selects one of them uniformly for replacement by F_i. The same draw therefore produces a paired human-only panel containing k human labels and a mixed panel containing $k-1$ human labels and one model label.

Let $\mathcal{L}$ be a finite label space and let $\delta(a,b)$ be a fixed, bounded, symmetric disagreement function satisfying $\delta(a,a)=0$ and $\delta(a,b)=\delta(b,a)$. For panel type $t \in \{h,m\}$, where h stands for human-only and m stands for mixed, let $Y_{t,1},\dots,Y_{t,k}$ denote its labels and define

$$d_t(Z,U)=\binom{k}{2}^{-1}\sum_{r<s}\delta(Y_{t,r},Y_{t,s}), \qquad p_t(a;Z,U)=\frac{1}{k}\sum_{r=1}^{k}\mathbf{1}\{Y_{t,r}=a\}.$$

Thus, $d_t(Z,U)$ is the average pairwise disagreement within the panel and $p_t(a;Z,U)$ is the proportion of labels assigned to category a.

The corresponding population quantities are

$$D_{o,t}=\mathbb{E}[d_t(Z,U)], \qquad P_t(a)=\mathbb{E}[p_t(a;Z,U)], \qquad D_{e,t}=\sum_{a,b\in\mathcal{L}}P_t(a)P_t(b)\delta(a,b), \qquad R_t=1-\frac{D_{o,t}}{D_{e,t}},$$

where the expectations are over both the item and panel draws.

The population substitution gap is

$$\Delta = R_m - R_h.$$

When panel size varies, we write $R_{m,k}$ and Δ_k to denote the panel size.

For N sampled panels, the empirical agreement coefficient is obtained by replacing $D_{o,t}$ and P_t with their pooled sample averages and applying the usual finite-sample correction to $D_{e,t}$. This is Krippendorff's coefficient for complete panels of equal size and the specified disagreement function (Krippendorff, 2011).

For a fixed panel size k, define

$$\gamma = c\frac{k_c}{k}, \qquad \varepsilon=\gamma(1-R_h), \qquad \Delta_{\mathrm{adj}}=\Delta+\varepsilon.$$

The non-inferiority hypotheses are

$$H_0:\Delta_{\mathrm{adj}}\le 0 \qquad \text{versus} \qquad H_1:\Delta_{\mathrm{adj}}>0.$$

In each bootstrap replicate, N items are sampled with replacement from the observed dataset. A fresh panel draw is generated independently for every sampled occurrence, and the corresponding human-only and mixed panels are constructed jointly. The bootstrap statistic is

$$\widehat{\Delta}^{b}_{\mathrm{adj}}=\widehat{R}^{b}_{m}-\widehat{R}^{b}_{h}+\gamma\big(1-\widehat{R}^{b}_{h}\big).$$

Thus, both agreement coefficients and the adaptive allowance are re-estimated within every replicate.

Let B_N denote the number of bootstrap replicates and define

$$q^{*}_{\alpha}=Q_{\alpha}\left(\left\{\widehat{\Delta}^{b}_{\mathrm{adj}}\right\}_{b=1}^{B_N}\right),$$

where Q_α is the empirical lower α-quantile. The LLM passes when $q^{*}_{\alpha}>0$.

B.2 Panel-Size Scaling

Let $P_F(a)=\mathbb{P}(F=a)$ denote the marginal distribution of the model labels. We further define the expected within-item model–human disagreement

$$D_{FH}=\mathbb{E}\left[\frac{1}{|H|}\sum_{y\in H}\delta(F,y)\right],$$

and the chance disagreements

$$G_{HF}=\sum_{a,b\in\mathcal{L}}P_h(a)P_F(b)\delta(a,b), \qquad G_F=\sum_{a,b\in\mathcal{L}}P_F(a)P_F(b)\delta(a,b).$$

Uniform sampling ensures that the marginal distribution of a selected human label and the joint distribution of a selected pair of distinct human labels do not depend on k. Consequently, P_h, $D_{o,h}$, $D_{e,h}$, and R_h remain fixed across feasible panel sizes.

A mixed panel contains $\binom{k-1}{2}$ human–human pairs and $k-1$ model–human pairs. Its observed disagreement is therefore

$$D_{o,m,k} = D_{o,h} + \frac{2}{k}(D_{FH} - D_{o,h}). \tag{3}$$

Its marginal label distribution is

$$P_{m,k}(a) = \frac{k-1}{k} P_h(a) + \frac{1}{k} P_F(a),$$

which yields

$$D_{e,m,k} = D_{e,h} + \frac{2}{k}(G_{HF} - D_{e,h}) + \frac{1}{k^2}(D_{e,h} - 2G_{HF} + G_F). \tag{4}$$

Because every term in the equivalent expression

$$D_{e,m,k} = \frac{(k-1)^2 D_{e,h} + 2(k-1)G_{HF} + G_F}{k^2}$$

is nonnegative,

$$D_{e,m,k} \geq \left(\frac{k-1}{k}\right)^2 D_{e,h}.$$

Thus, $D_{e,h} > 0$ also guarantees $D_{e,m,k} > 0$.

Proposition 1 (First-order panel-size scaling)**.** *Suppose that $D_{o,h}$, $D_{e,h}$, D_{FH}, G_{HF}, and G_F remain fixed as k varies and that $D_{e,h} > 0$. Then*

$$\Delta_k = \frac{2(D_{o,h}G_{HF} - D_{FH}D_{e,h})}{D_{e,h}^2} \frac{1}{k} + O(k^{-2}). \tag{5}$$

Hence, $\Delta_k = O(k^{-1})$. If the human and model marginal label distributions coincide, then

$$\Delta_k = \frac{2(D_{o,h} - D_{FH})}{D_{e,h}} \frac{1}{k} \tag{6}$$

holds exactly for every admissible panel size k.

Proof. Equations 3 and 4 imply that the difference in observed disagreements between the mixed and human-only panels is

$$D_{o,m,k} - D_{o,h} = \frac{2}{k}(D_{FH} - D_{o,h})$$

and the expected difference in disagreements is

$$D_{e,m,k} - D_{e,h} = \frac{2}{k}(G_{HF} - D_{e,h}) + O(k^{-2}).$$

Expanding

$$\Delta_k = R_{m,k} - R_h = -\frac{D_{o,m,k}}{D_{e,m,k}} + \frac{D_{o,h}}{D_{e,h}}$$

around $(D_{o,h}, D_{e,h})$ gives

$$\Delta_k = -\frac{D_{o,m,k} - D_{o,h}}{D_{e,h}} + \frac{D_{o,h}}{D_{e,h}^2}(D_{e,m,k} - D_{e,h}) + O(k^{-2}).$$

Substituting the two preceding expressions and collecting the terms of order $1/k$ proves Equation 5.

If $P_F = P_h$, then $G_{HF} = G_F = D_{e,h}$, and Equation 4 gives $D_{e,m,k} = D_{e,h}$. Substitution of Equation 3 then yields Equation 6. □

Proposition 1 shows that replacing one member of a larger panel produces a proportionally smaller change in agreement. This motivates scaling the allowance at the same rate. For a reference panel size $k_c \geq 3$, we therefore define

$$\varepsilon_k = c(1 - R_h)\frac{k_c}{k}.$$

This adjustment is exact when the human and model label marginals coincide and provides a first-order normalization otherwise. The proposition concerns variation in k across a family of admissible panel constructions. The asymptotic result below instead fixes k and lets the number of items N increase.

B.3 Consistency and Asymptotic level α

The proof relies on a first-order expansion of the agreement coefficient. For panel type t, define its influence contribution as

$$\ell_t(Z,U) = -\frac{d_t(Z,U) - D_{o,t}}{D_{e,t}} + \frac{2D_{o,t}}{D_{e,t}^2} \sum_{a,b\in\mathcal{L}} P_t(b)\delta(a,b)\big[p_t(a;Z,U) - P_t(a)\big]. \tag{7}$$

The first term represents the panel's contribution to observed disagreement, and the second represents its contribution through the marginal label distribution and chance disagreement. Equation 7 is the influence function obtained by differentiating $R_t = 1 - D_{o,t}/D_{e,t}$ with respect to the distribution of panels.

The adjusted substitution gap has influence contribution

$$\ell(Z,U) = \ell_m(Z,U) - (1+\gamma)\ell_h(Z,U).$$

Let

$$\bar{\ell}(Z) = \mathbb{E}[\ell(Z,U) \mid Z], \qquad \sigma^2 = \mathrm{Var}[\bar{\ell}(Z)], \qquad \tau^2 = \mathrm{Var}[\ell(Z,U)].$$

Lemma 1 (Variance contributed by panel selection)**.** *If $D_{e,h} > 0$, then ℓ and $\bar{\ell}$ are bounded and have mean zero. Moreover,*

$$\tau^2 = \sigma^2 + \mathbb{E}\left[\mathrm{Var}[\ell(Z,U) \mid Z]\right] \geq \sigma^2. \tag{8}$$

Proof. Boundedness follows from the bounded disagreement function, the finite label space, and the positivity of $D_{e,h}$ and $D_{e,m}$. The centered terms in Equation 7 have expectation zero, implying $\mathbb{E}[\ell] = \mathbb{E}[\bar{\ell}] = 0$. Equation 8 follows directly from the law of total variance conditional on Z. Because the human-only and mixed contributions use the same item and panel draw, their paired dependence is retained in ℓ. □

The distinction between σ^2 and τ^2 is important. The sampling variance σ^2 reflects variation across item clusters after averaging over the possible panel draws for each item. The bootstrap variance τ^2 additionally includes the variation introduced by drawing a fresh panel for every bootstrap occurrence. The latter is therefore at least as large as the former, making the resulting lower bound potentially conservative.

Theorem 1 (Consistency and Asymptotic level α)**.** *Consider a fixed population distribution in which the item clusters $Z_1, \ldots, Z_N$ are independent and identically distributed. Assume that the label space, bounded disagreement function, and panel size $k \geq 3$ remain fixed as $N \to \infty$, and that every item has at least k human annotations. Suppose that $D_{e,h} \geq d_0$ for some constant $d_0 > 0$, while the model, prompt, c, k_c, and significance level $0 < \alpha < 1/2$ are fixed independently of the test sample. Assume further that $\sigma^2 > 0$, that the bootstrap replicates are conditionally independent and use fresh independent panel draws, and that $B_N \to \infty$.*

Let Φ denote the standard normal cumulative distribution function, let $z_\alpha = \Phi^{-1}(\alpha)$, and let σ and τ be the positive square roots of the variances defined above. Then

$$q_\alpha^* \xrightarrow{p} \Delta_{\mathrm{adj}},$$

and

$$\mathbb{P}\{\Delta_{\mathrm{adj}} \geq q_\alpha^*\} \longrightarrow \Phi\Big(-z_\alpha \frac{\tau}{\sigma}\Big) \geq 1 - \alpha.$$

Consequently,

$$\begin{aligned}
\Delta_{\mathrm{adj}} \leq 0 \quad &\Longrightarrow \quad \limsup_{N\to\infty} \mathbb{P}\{q_\alpha^* > 0\} \leq \alpha, \\
\Delta_{\mathrm{adj}} < 0 \quad &\Longrightarrow \quad \mathbb{P}\{q_\alpha^* > 0\} \longrightarrow 0, \\
\Delta_{\mathrm{adj}} > 0 \quad &\Longrightarrow \quad \mathbb{P}\{q_\alpha^* > 0\} \longrightarrow 1.
\end{aligned}$$

All probabilities account jointly for item sampling and bootstrap randomization.

Proof. For each observed item, first average its panel disagreement and label proportions over the conditional distribution of U. Substituting these conditional averages into the pooled agreement formula defines an item-level estimator $\widehat{\Delta}_{\mathrm{adj},N}$. This quantity is used only in the proof; the test does not need to compute it.

Because its inputs are averages of bounded i.i.d. item-level quantities, the law of large numbers gives

$$\widehat{\Delta}_{\mathrm{adj},N} \xrightarrow{p} \Delta_{\mathrm{adj}}.$$

The agreement coefficient is a smooth function of these averages whenever its chance-disagreement denominator is bounded away from zero. The multivariate delta method and Equation 7 therefore yield

$$\sqrt{N}\big(\widehat{\Delta}_{\mathrm{adj},N} - \Delta_{\mathrm{adj}}\big) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\ell}(Z_i) + o_p(1) \Longrightarrow \mathcal{N}(0,\sigma^2). \tag{9}$$

The finite-sample correction in the empirical chance disagreement contributes only $O(N^{-1})$ and therefore does not affect this limit.

For a generic bootstrap replicate, let I_j be the item index sampled at occurrence j and let U_j be its fresh panel draw. Conditional on the observed items, the pairs (Z_{I_j}, U_j) are i.i.d. Marking with a star the conditioning on the observed items and applying the same delta-method expansion gives

$$\sqrt{N}\big(\widehat{\Delta}^*_{\mathrm{adj}} - \widehat{\Delta}_{\mathrm{adj},N}\big) \Longrightarrow^* \mathcal{N}(0, \tau^2) \quad \text{in probability.} \tag{10}$$

Formally, the bootstrap expansion first uses the empirical, plug-in version of Equation 7. Its coefficients converge in probability to their population counterparts because the empirical disagreements and label proportions are consistent. Boundedness then makes the difference between the empirical and population influence contributions $o_{p^*}(1)$ on the scale in Equation 10.

The conditional distribution in Equation 10 has asymptotic lower α-quantile τz_α. Because $B_N \to \infty$, the empirical bootstrap quantile converges to the corresponding conditional quantile. Hence,

$$q^*_\alpha = \widehat{\Delta}_{\mathrm{adj},N} + \frac{\tau z_\alpha}{\sqrt{N}} + o_p(N^{-1/2}). \tag{11}$$

Consistency of q^*_α follows immediately.

Combining Equations 9 and 11 yields

$$\sqrt{N}\big(q^*_\alpha - \Delta_{\mathrm{adj}}\big) \Longrightarrow \mathcal{N}(\tau z_\alpha, \sigma^2).$$

Therefore,

$$\mathbb{P}\{\Delta_{\mathrm{adj}} \geq q^*_\alpha\} \longrightarrow \Phi\Big(-z_\alpha \frac{\tau}{\sigma}\Big).$$

Since $z_\alpha < 0$ and $\tau \geq \sigma$ by Lemma 1, this limit is at least $1 - \alpha$.

Under H_0, the event $\{q^*_\alpha > 0\}$ is contained in $\{q^*_\alpha > \Delta_{\mathrm{adj}}\}$, so the asymptotic false-pass probability is at most α. If $\Delta_{\mathrm{adj}} < 0$ or $\Delta_{\mathrm{adj}} > 0$, the remaining conclusions follow from $q^*_\alpha \xrightarrow{p} \Delta_{\mathrm{adj}}$. □

C Simulations: Additional Results

All results use $c = 0.1$ unless stated otherwise.

Comparison with the baselines for different item sizes. Figure 3 repeats the comparison at $N \in \{50, 100, 150, 300\}$ on identical datasets and subsamples. The gap is largest with few items: at the worse-model pass rate the ALT-TEST obtains with $\varepsilon = 0.10$, the ALTRA-TEST reaches better-model pass rates of 0.855, 0.927, and 0.987 at $N = 50$, 100, and 150, against 0.677, 0.898, and 0.958. As N grows, both procedures approach 1.00, so we expect the difference to vanish asymptotically. The recommended settings differ more: the ALT-TEST with $\varepsilon = 0.10$ produces worse-model pass rates of 0.033, 0.021, 0.049, and 0.088, with condition-level rates up to 0.24, and its margins for more skilled annotators, $\varepsilon = 0.15$ and 0.20, remain outside the admissible region at every size, reaching 0.41 and 0.62 at $N = 300$, whereas the ALTRA-TEST with $c = 0.1$ stays between 0.003 and 0.010, with no condition above 0.08. STT-ALT approaches the other frontiers for $N \geq 100$, reaching 0.908, 0.911, and 0.960 at its smallest scaling factor, but has no admissible setting at $N = 50$, where even $\mathrm{frac} = 0.01$ gives a worse-model pass rate of 0.081; this rate falls to 0.032, 0.019, and 0.003 as N grows, consistent with the STT treating bootstrap replicates as independent observations and thereby underestimating sampling variability.

Generative family. Across the three families defined in Appendix A.3, the worse-model pass rate ranges from 0.009 to 0.033 and the better-model pass rate is at least 0.984. A fixed margin does not transfer in the same way: the ALT-TEST is admissible at $\varepsilon = 0.075$ but not at $\varepsilon = 0.10$ under the Dirichlet family, and admissible at $\varepsilon = 0.125$ under the uniform and Gaussian families. Figure 4 shows that both alternative families are easier than the Dirichlet baseline, with the ALTRA-TEST and ALT-TEST attaining better-model pass rates of 1.00 within the admissible region; the ALTRA-TEST frontier still lies at or above the ALT-TEST frontier at nearly every admissible setting and reaches 1.00 at a lower worse-model pass rate, while STT-ALT performs below both. Since STT-ALT tests the same non-inferiority hypothesis, the remaining difference is procedural: it compares the average agreement of substituted panels with the all-human agreement and estimates its margin from two disjoint human groups, whereas the ALTRA-TEST pairs every mixed panel with its own control. The Dirichlet family used in Section 4 is therefore the most difficult of the three and gives the most conservative comparison.

Equivalence. Replacing only the decision rule of the ALTRA-TEST by the equivalence rule of Appendix A.1, and holding the datasets, subsamples, bootstrap, coefficient, allowance and significance level fixed, reduces the better-model pass rate from 0.984 to 0.030 at $c = 0.1$; the percentile form used here disagrees with equivalent t- and z-based implementations on 0.68% of the 2500 runs. This criterion is not the right one for an equivalence test, however, since it treats acceptance of a better-than-human model as a success, whereas equivalence rejects large differences in either direction. Figure 5 (left) therefore reports the pass rate when equal, across the five diagonal conditions with $\eta_f = \eta_h$, against the pass rate when not equal, across the twelve off-diagonal conditions with $|\eta_f - \eta_h| \geq 0.2$. Under these definitions, the EQ-ALTRA-TEST attains higher pass rates when equal than the STT within the admissible region, and the two frontiers are nearly indistinguishable elsewhere. Figure 5 (right) fixes $c = 0.125$ and varies N: the pass rate when not equal stays below $\alpha = 0.05$ at every sample size, while the pass rate when equal rises from 0.07 at $N = 100$ to 0.36 at $N = 200$ and 0.52 at $N = 300$, so the equivalence variant needs substantially more items to pass models of equal quality than the non-inferiority test needs to pass better ones.

Calibration of the allowance constant. Figure 6a presents the better-model and worse-model pass rates across values of c. The better-model pass rate increases sharply until $c = 0.1$ and changes little thereafter, whereas the worse-model pass rate remains below 0.01 through $c = 0.1$ and then rises, reaching $\alpha = 0.05$ near $c = 0.135$; the variability across conditions follows the same pattern, with the worse-model band widening beyond $c = 0.1$. At $c = 0.1$ the two rates are 0.984 and 0.009, and larger values add less than two percentage points to the better-model pass rate while leaving little margin for error control under harder conditions. Figure 6b reports the pass rate for each pair (η_h, η_f) at that setting: no better-model condition falls below 0.92, and the only worse-model condition above α is $(\eta_h, \eta_f) = (0.4, 0.5)$ at 0.08, where the model's deficit still lies within the allowance ($\Delta/\varepsilon = -0.54$), so passing it is consistent with the formal non-inferiority hypothesis. Across the nine conditions that fall within that hypothesis, the mean pass rate is 0.001 and no condition exceeds 0.01.

The three structural parameters are swept separately, each repeating the complete noise grid at $N = 200$ and $N = 100$ with the remaining parameters at their default values.

Group size k. Varying k from 3 to 7 with the allowance scaled by k_c/k, the better-model pass rate remains between 0.979 and 0.995 at $N = 200$ and the mean worse-model pass rate between 0.008 and 0.025 (Figure 7a); at $N = 100$, the better-model pass rate is lower at $k = 3$ and stable for $k \geq 4$ (Figure 7b). This stability is consistent with the panel-size adjustment derived in Appendix B.2: without it, replacing one annotation would affect agreement less as the panel grows, while the permitted margin stayed fixed.

Annotators per item m. Increasing m from 3 to 12 moves the better-model pass rate from 0.962 to 0.988 at $N = 200$ and the mean worse-model pass rate from 0.022 to 0.006 (Figure 8a); at $N = 100$, the better-model pass rate is lower and increases more noticeably with m (Figure 8b). Because only k labels are sampled per item, a larger pool mainly reduces variability in panel construction, so small annotator pools benefit from more items.

Number of labels L. Binary tasks are the most difficult setting, with a better-model pass rate of 0.823 at $N = 200$, compared with at least 0.98 for $L \geq 4$ (Figure 9a), and lower still at $N = 100$ (Figure 9b). With two categories, chance agreement is 0.5, so a given error rate induces a smaller change in the chance-corrected coefficient and the substitution gap is harder to detect. Binary tasks therefore require more items to reach the same better-model pass rate.

Agreement metric. Figure 10 repeats the noise grid at $N = 200$ with four alternative agreement formulations. Krippendorff's α attains the highest better-model pass rate with a worse-model pass rate below 0.01, and Fleiss' κ is nearly identical. Percent agreement is slightly lower, since it is bounded on the same scale but does not correct for chance and responds less sharply to substitutions that alter the marginal label distribution. The error-based measures are computed on the label indices and mapped to agreement scales as $R = 1 - \mathrm{MSE}/(L-1)^2$ and $R = 1 - \mathrm{RMSE}/(L-1)$: MSE keeps the worse-model pass rate low but reduces the better-model pass rate to 0.88, and RMSE reaches a worse-model pass rate of 0.088. Neither transformation corrects for chance, so the same c corresponds to a different tolerance depending on the metric's scaling and no longer reads as a fixed proportion of the human disagreement.

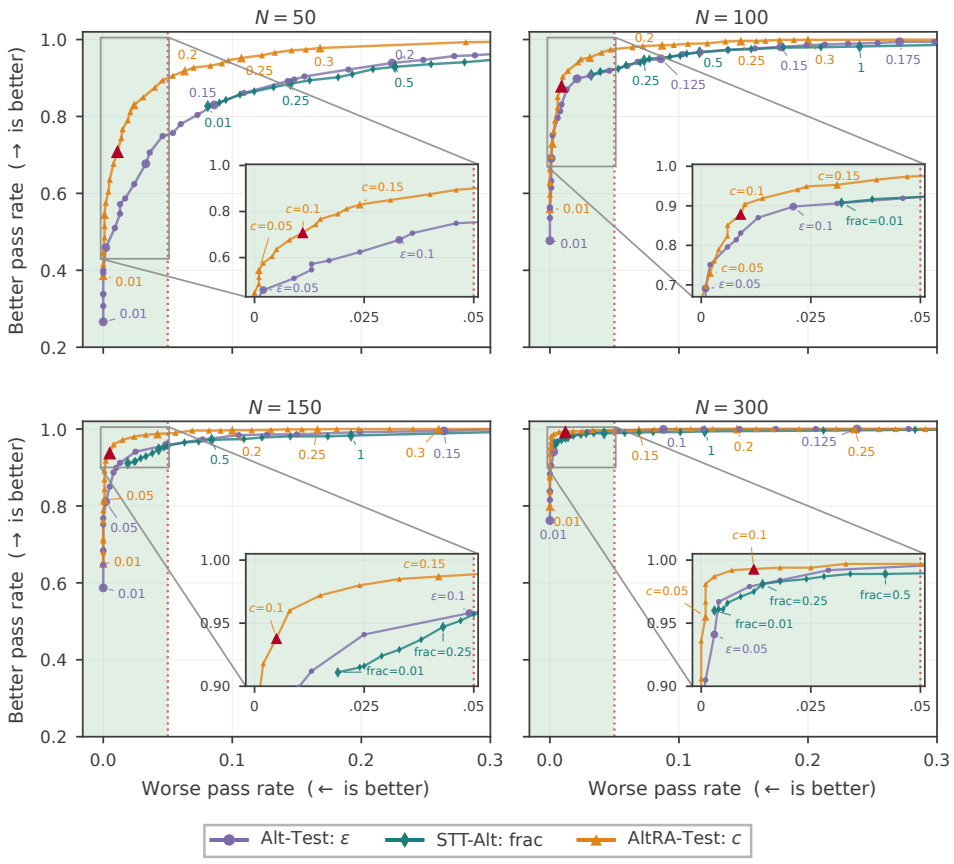
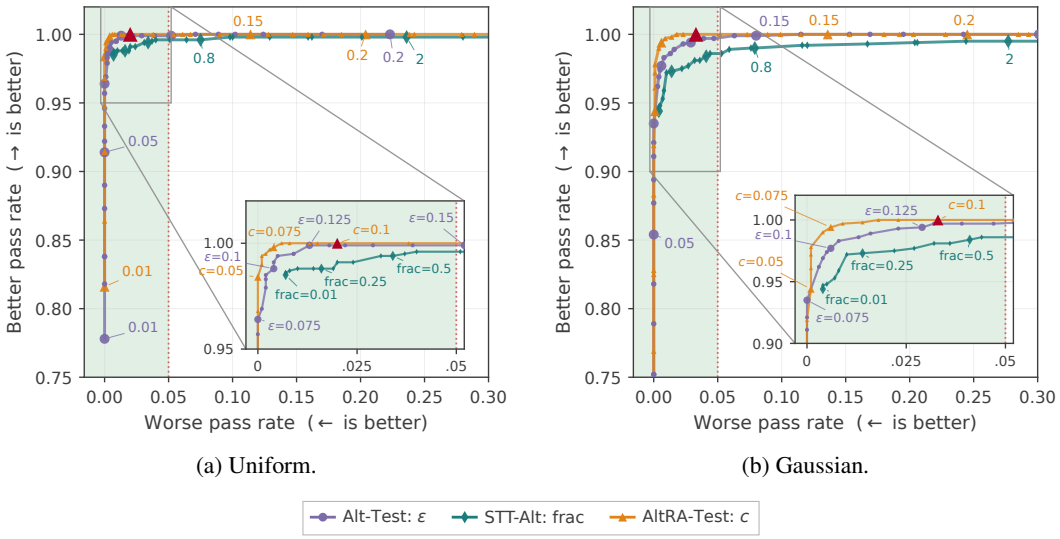


Figure 3: Comparison using simulations with $k = 3$ at $N = 50$, 100, 150, and 300 items, while varying the hyperparameter of each procedure: c, ε, or frac. The AltRA-Test dominates the baselines at every N, with the largest gains at smaller N.



(a) Uniform. (b) Gaussian.

Figure 4: Simulations with different data-generating distribution families: uniform on the left and Gaussian on the right. As shown, the AltRA-Test dominates the baselines even when changing distributions.

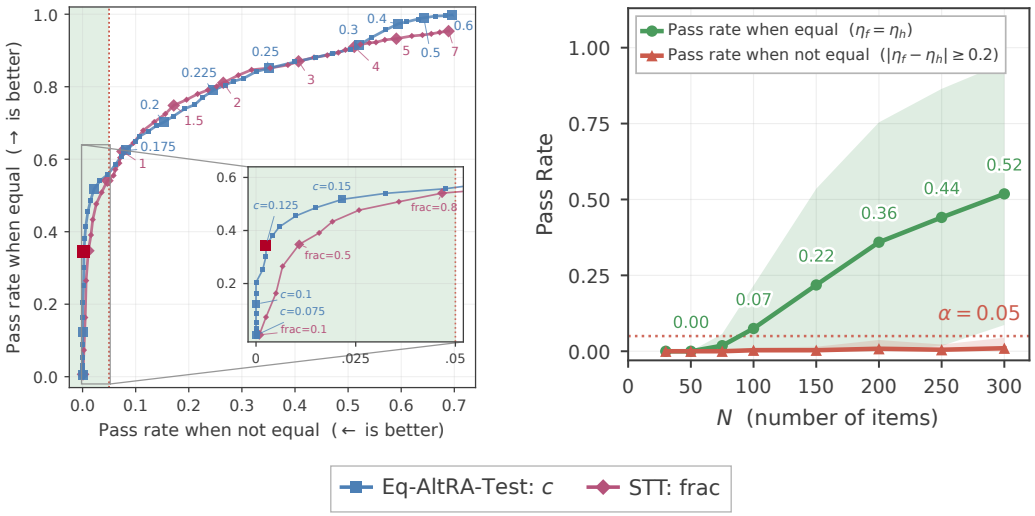
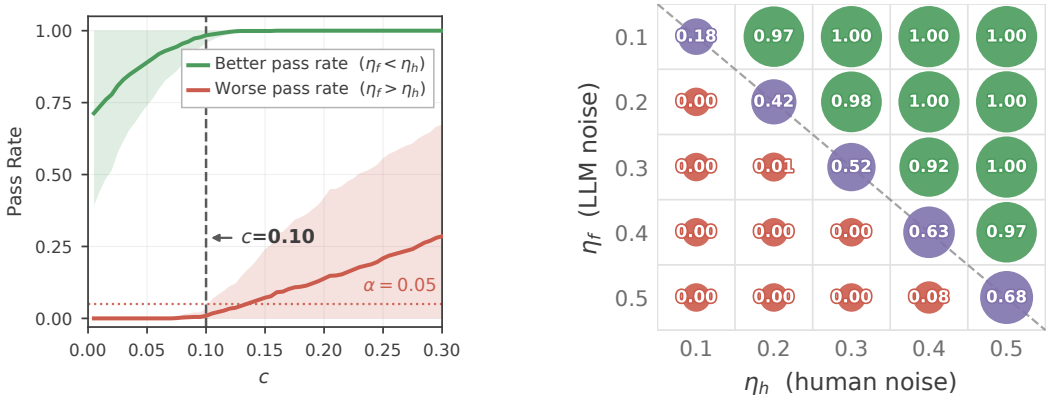


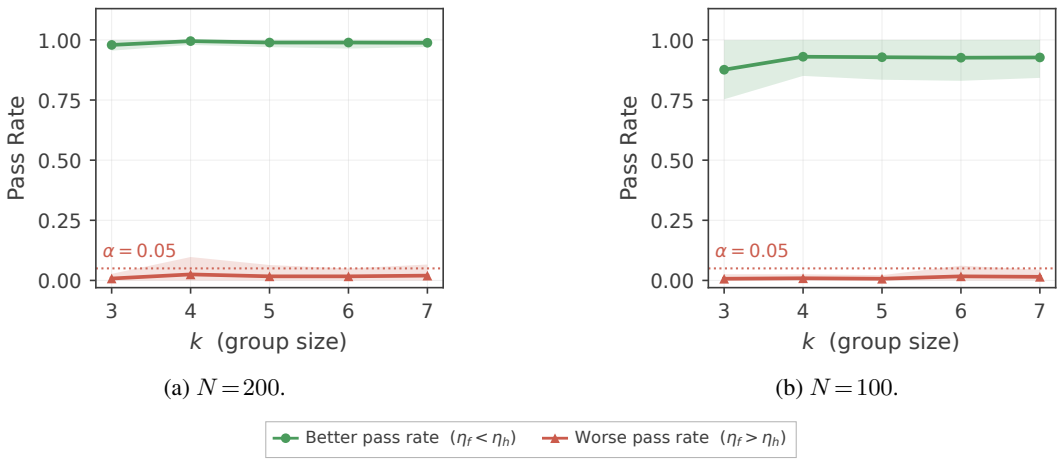
Figure 5: *Left:* Comparison of the EQ-ALTRA-TEST (Appendix A.1) against the STT for the equivalence question. Both apply two one-sided tests: the EQ-ALTRA-TEST to bootstrap quantiles, and the STT as t-tests. Simulations use $N = 200$ items and $k = 3$, while varying c for the EQ-ALTRA-TEST and frac for the STT. The X-axis is the pass rate when not equal, over conditions with $|\eta_f - \eta_h| \geq 0.2$, and the Y-axis is the pass rate when equal, over conditions with $\eta_f = \eta_h$. *Right:* Using the calibration in the left panel, we set $c = 0.125$ and vary the number of items N.



(a) Better-model and worse-model pass rates against c.

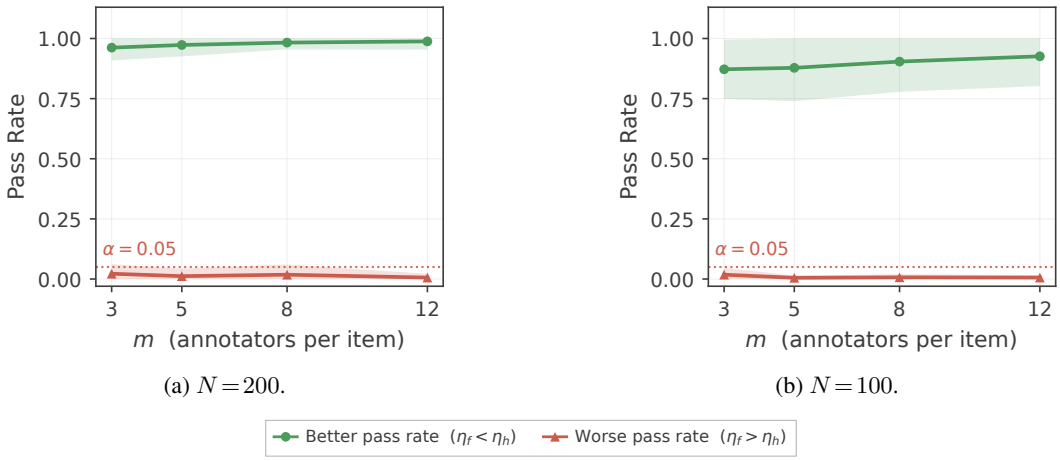
(b) Pass rate over the noise grid at $c = 0.1$.

Figure 6: Calibration and behavior of the ALTRA-TEST. (a) Each curve is the mean over the conditions of its regime and the shaded band is one standard deviation across them; the dashed line marks the recommended $c = 0.1$. (b) Per-condition pass rate at that setting; passing is correct where the LLM is less noisy, an error where it is noisier, and neither on the diagonal.



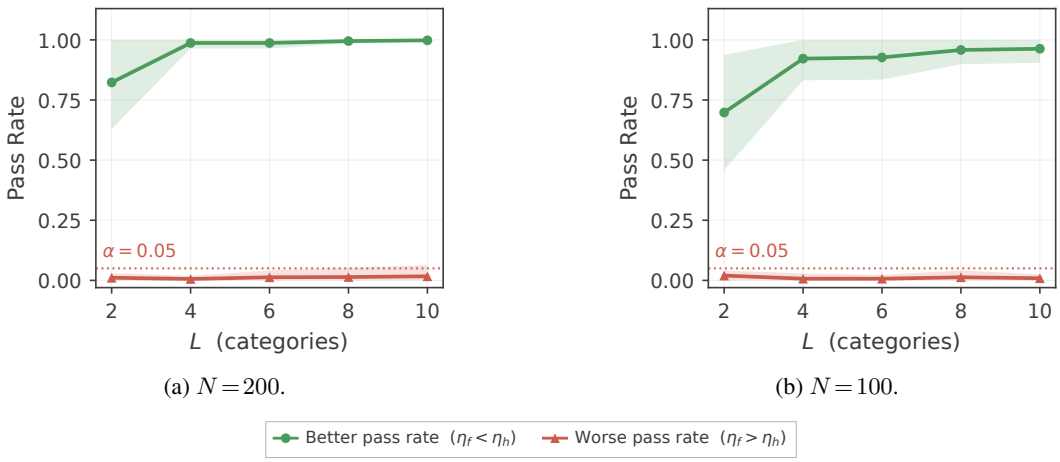
(a) $N = 200$. (b) $N = 100$.

Figure 7: Group size k, with the margin scaled by k_c/k, at $c = 0.1$. The worse-model pass rate stays below 0.05 throughout.



(a) $N = 200$. (b) $N = 100$.

Figure 8: Annotators per item m, at $c = 0.1$. The worse-model pass rate stays below 0.05 throughout.



(a) $N = 200$. (b) $N = 100$.

Figure 9: Number of labels L, at $c = 0.1$. The worse-model pass rate stays below 0.05 throughout.

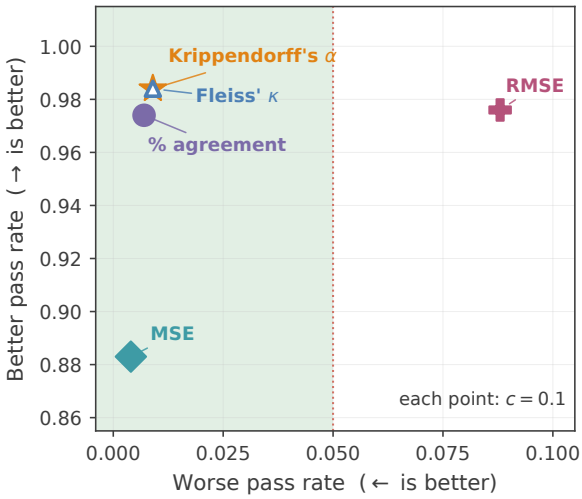


Figure 10: Simulations of the ALTRA-TEST with $c = 0.1$ for different IAA metrics.

D REAL-WORLD DATA: ADDITIONAL RESULTS

This appendix reports additional results for the real-world experiments of Section 5: the per-dataset outcomes for every judge and prompting technique, a comparison with the ALT-TEST and the STT on the datasets that support them, and the sensitivity of the verdicts to the number of items, the number of few-shot examples, and the CoT reasoning effort.

Per-dataset results. Tables 6–9 report the complete results underlying the verdicts in Section 5, including Measuring-Hate-Speech and the judges omitted from Tables 1 and 2. Each block places two datasets side by side, with one row per judge and one column per prompting technique. Each cell reports $\bar{R}_m$, and its color indicates whether the judge passes the test, i.e., whether $q^*_\alpha > 0$; each block also reports $\bar{R}_h$, the baseline against which $\bar{R}_m$ should be read. Gray crosses mark configurations that were not evaluated.

Comparison with existing tests. Table 10 compares the procedures on the datasets with persistent annotator identities, with each judge at its best prompting technique and the ALT-TEST at $\varepsilon = 0.1$; since all tests use the items valid for every test. Over the 94 judge–dataset pairs they share, the ALT-TEST and the ALTRA-TEST agree on 68 verdicts and pass 42 and 44, both passing every judge on CEBaB and none on Loneliness. Their disagreements are dataset-specific rather than directional: the ALTRA-TEST is stricter on four datasets, more permissive on three, and identical on two, with the largest gaps on EPIC (0 versus 9 passes), DICES-350 (9 versus 4), SummEval (6 versus 2), and 10k-Prompts (5 versus 8). On DICES-350, the STT passes only Gem-P and GPT-L and returns $p_{\mathrm{TOST}} = 1.000$ for six judges, since equivalence also fails a judge that raises agreement; STT-ALT additionally passes G3-12b ($p_{\mathrm{lower}} < 0.01$), so its remaining gap to the four ALTRA-TEST passes reflects margins and estimation rather than the hypothesis.

Sample size. Figures 11 and 12 report q^*_α as a function of the number of evaluated items for all 20 datasets, with one curve per judge at its best prompting technique; the dashed line marks the smallest N at which any judge first passes. We call a judge's curve stable at the smallest N from which it stays within 0.02 of its value at $N = 300$, and denote by $\bar{N}$ the mean stabilization point across a dataset's judges. Seven datasets stabilize well before $N = 200$ (FER+ at $\bar{N} = 122$, CEBaB at 125, CIFAR-10H and Flickr8k-Expert at 135, Scruples at 157, ChaosNLI at 173, SOMOS at 186), five near it (T2V-Preference and Vibe-Arena at 190, ConvAbuse at 198, EPIC at 205, Measuring-Hate-Speech at 207), and eight after it (DICES-350 and SummEval at 214, T2I-DPO at 220, TTS-Preference at 221, Persuasion at 223, 10k-Prompts at 234, Loneliness at 236, MultiPref at 239). Even on these, no judge changes by more than 0.05 between $N = 200$ and $N = 300$, and on seven datasets every judge has stabilized by $N = 200$. Stabilization tracks task difficulty and label distribution more than modality, arriving earlier on high-agreement tasks and later on low-agreement or long-tailed ones. The verdict settles earlier than the value of q^*_α: judges that pass with 300 items typically already pass with about 100 (Table 2). We therefore recommend at least 100 items, consistent with the simulations in Section 4; more items mainly make q^*_α, and hence the comparison between judges, more precise.

Number of few-shot examples. Figure 13 reports q^*_α as the number of randomly sampled few-shot examples varies from zero to fifteen, on five datasets spanning text, image, and audio. On CEBaB, the curves

are nearly identical once at least three examples are used, and on ChaosNLI most of the change occurs by three examples. EPIC is more sensitive and not monotonic: L3.1-8b drops sharply at three examples before partially recovering at fifteen. T2I-DPO largely stabilizes by five examples, whereas on SOMOS, Gem-P and Gem-L continue to improve up to fifteen. Only four judge–dataset verdicts change beyond five examples, and all but Gem-L on SOMOS have decision statistics close to zero. We therefore use five examples in the main experiments, which captures most of the observed benefit while limiting context length and computation.

CoT reasoning effort. Figure 14 reports q^*_α at low (zero), mild, and high CoT reasoning effort for the three Gemini judges, whose effort is set by the API thinking-effort parameter, on five datasets spanning text, image, and audio. Across the fifteen curves, the mean decision statistic increases from 0.017 at low effort to 0.023 at mild and 0.036 at high effort, and high effort outperforms low effort on ten of the fifteen. The gains are concentrated in a few cases, however, and several curves move in the opposite direction, so no level dominates across datasets. We use mild effort in the main experiments, which captures part of the average improvement at substantially lower cost than high effort. This also indicates that the low aggregate pass count of CoT in Section 5.2 cannot be attributed solely to an insufficient reasoning budget.

Table 6: **Text.** Test outcomes and mixed-panel agreement $\bar{R}_m$ for every text judge and prompting technique. Dark green marks the judge's best technique when it passes; light green marks a passing technique that is not the judge's best; gray marks the judge's best technique when it fails. **Bold** marks the highest $\bar{R}_m$ for each dataset.

Family	Judge	Zero	Few	CoT	Ens	GEPA	Zero	Few	CoT	Ens	GEPA
		DICES-350 ($\bar{R}_h = 0.20$)					**EPIC** ($\bar{R}_h = 0.20$)				
Gemini	Gem-P	0.14	0.20	0.14	0.15	✗	0.21	0.23	0.22	0.22	✗
	Gem-F	0.05	0.10	0.06	0.05	✗	0.19	0.16	0.19	0.19	✗
	Gem-L	0.10	0.12	0.10	0.10	✗	0.20	0.20	0.24	0.22	✗
GPT	GPT-T	0.14	0.19	0.16	0.17	✗	0.22	0.23	0.22	0.23	✗
	GPT-L	0.16	0.19	0.16	0.18	✗	0.23	**0.28**	0.21	0.24	✗
Qwen	Q3-8b	0.06	0.06	0.09	0.07	0.06	0.19	0.26	0.20	0.26	0.24
	Q3-14b	0.15	0.16	0.16	0.16	0.15	0.24	0.25	0.22	0.25	0.22
Gemma	G3-4b	0.11	0.12	0.12	0.12	0.11	−0.01	0.09	−0.06	0.09	−0.02
	G3-12b	0.19	0.18	**0.22**	0.19	0.19	0.22	0.24	0.19	0.24	0.24
Llama	L3.2-3b	0.09	0.10	0.14	0.11	0.09	0.13	0.16	0.01	0.19	−0.14
	L3.1-8b	0.09	0.14	0.15	0.15	0.09	0.09	−0.02	0.02	0.06	0.09
		MultiPref ($\bar{R}_h = 0.28$)					**ChaosNLI** ($\bar{R}_h = 0.28$)				
Gemini	Gem-P	**0.33**	0.31	**0.33**	**0.33**	✗	0.31	0.30	0.31	0.31	✗
	Gem-F	0.32	0.27	0.30	0.32	✗	0.28	0.28	0.30	0.29	✗
	Gem-L	0.32	0.26	0.28	0.31	✗	0.32	0.30	0.29	0.31	✗
GPT	GPT-T	0.25	0.24	0.24	0.25	✗	**0.33**	0.26	0.31	**0.33**	✗
	GPT-L	0.24	0.24	0.24	0.25	✗	0.29	0.32	0.26	0.30	✗
Qwen	Q3-8b	0.18	0.09	0.19	0.14	0.18	0.26	0.30	0.26	0.28	0.26
	Q3-14b	0.28	0.19	0.26	0.30	0.26	0.32	0.32	0.32	0.32	0.31
Gemma	G3-4b	0.02	0.16	0.14	0.24	0.02	0.27	0.29	0.28	0.29	0.27
	G3-12b	0.25	0.21	0.25	0.26	0.25	0.31	**0.33**	0.29	0.32	0.30
Llama	L3.2-3b	0.04	0.02	0.20	0.03	0.01	0.18	0.20	0.23	0.20	0.18
	L3.1-8b	0.12	0.16	0.12	0.19	0.11	0.28	0.29	0.28	0.29	0.28
		Persuasion ($\bar{R}_h = 0.31$)					**10k-Prompts** ($\bar{R}_h = 0.37$)				
Gemini	Gem-P	0.07	0.06	0.06	0.06	✗	0.31	0.44	0.40	0.43	✗
	Gem-F	0.09	0.17	0.07	0.08	✗	0.36	0.37	0.36	0.42	✗
	Gem-L	0.01	0.05	0.01	0.02	✗	**0.45**	0.41	0.43	**0.45**	✗
GPT	GPT-T	0.18	**0.27**	0.15	0.17	✗	0.38	0.38	0.35	0.38	✗
	GPT-L	0.08	0.17	0.10	0.10	✗	0.38	0.35	0.35	0.40	✗
Qwen	Q3-8b	0.24	0.24	0.12	0.25	0.24	0.39	0.38	0.39	0.42	0.40
	Q3-14b	0.19	0.22	0.18	0.23	0.19	0.39	0.39	0.15	0.41	0.41

Text (continued)

Family	Judge	Zero	Few	CoT	Ens	GEPA	Zero	Few	CoT	Ens	GEPA
Gemma	G3-4b	0.17	−0.01	0.18	0.11	0.17	0.36	0.12	0.33	0.23	0.36
	G3-12b	0.15	0.22	0.17	0.21	0.15	0.35	0.32	0.39	0.38	0.35
Llama	L3.2-3b	0.06	0.20	0.10	0.17	0.06	0.33	0.23	0.23	0.32	0.32
	L3.1-8b	0.06	0.06	0.09	0.05	0.06	0.27	0.28	0.28	0.32	0.27
		Scruples ($\bar{R}_h = 0.38$)					**Meas.-Hate-Speech** ($\bar{R}_h = 0.53$)				
Gemini	Gem-P	**0.45**	0.44	**0.45**	**0.45**	✗	0.50	**0.51**	0.50	0.50	✗
	Gem-F	0.44	0.43	0.44	**0.45**	✗	0.49	0.50	0.50	0.49	✗
	Gem-L	0.41	0.40	0.43	0.42	✗	0.47	0.47	0.45	0.46	✗
GPT	GPT-T	0.39	0.37	0.39	0.39	✗	0.47	0.44	0.48	0.47	✗
	GPT-L	0.36	0.32	0.35	0.36	✗	0.41	0.41	0.42	0.42	✗
Qwen	Q3-8b	0.08	0.15	0.18	0.10	0.08	0.29	0.43	0.31	0.36	0.35
	Q3-14b	0.04	0.14	0.10	0.06	0.08	0.29	0.40	0.30	0.34	0.35
Gemma	G3-4b	0.06	0.07	0.14	0.07	0.06	0.25	0.41	0.27	0.30	0.30
	G3-12b	0.24	0.13	0.27	0.25	0.22	0.38	0.46	0.45	0.45	0.38
Llama	L3.2-3b	−0.00	0.04	0.08	−0.02	−0.01	0.30	0.38	0.29	0.34	0.30
	L3.1-8b	−0.01	−0.00	0.23	−0.00	−0.01	0.28	0.36	0.41	0.38	0.28
		SummEval ($\bar{R}_h = 0.60$)					**ConvAbuse** ($\bar{R}_h = 0.72$)				
Gemini	Gem-P	0.64	0.62	**0.65**	**0.65**	✗	0.74	**0.75**	0.73	**0.75**	✗
	Gem-F	0.58	0.57	0.61	0.59	✗	0.72	0.73	0.71	0.71	✗
	Gem-L	0.57	0.53	0.58	0.58	✗	0.74	0.72	0.73	0.73	✗
GPT	GPT-T	0.56	0.51	0.54	0.55	✗	0.71	0.68	0.68	0.70	✗
	GPT-L	0.53	0.50	0.52	0.53	✗	0.73	0.74	0.71	**0.75**	✗
Qwen	Q3-8b	0.46	0.50	0.51	0.51	0.46	0.71	0.70	0.70	0.73	0.71
	Q3-14b	0.43	0.56	0.42	0.48	0.43	0.60	0.68	0.65	0.65	0.70
Gemma	G3-4b	0.37	0.27	0.42	0.34	0.37	0.24	0.44	0.41	0.35	0.25
	G3-12b	0.50	0.50	0.45	0.56	0.50	0.61	0.64	0.61	0.62	0.57
Llama	L3.2-3b	0.33	0.20	0.40	0.30	0.33	0.44	0.51	0.48	0.53	0.46
	L3.1-8b	0.50	0.38	0.28	0.49	0.50	0.43	0.61	0.58	0.58	0.09
		CEBaB ($\bar{R}_h = 0.72$)					**Loneliness** ($\bar{R}_h = 0.90$)				
Gemini	Gem-P	0.77	0.79	0.79	0.79	✗	0.57	**0.69**	0.57	0.59	✗
	Gem-F	0.73	0.79	0.79	0.79	✗	0.56	0.65	0.61	0.60	✗
	Gem-L	0.79	**0.80**	0.79	**0.80**	✗	0.42	0.58	0.43	0.47	✗
GPT	GPT-T	**0.80**	0.79	0.79	0.79	✗	0.32	0.59	0.38	0.39	✗
	GPT-L	0.79	0.79	0.79	0.79	✗	0.41	0.61	0.44	0.46	✗
Qwen	Q3-8b	0.78	0.77	0.78	0.78	0.78	0.28	0.46	0.18	0.38	0.37
	Q3-14b	0.78	0.78	0.77	0.78	0.79	0.34	0.47	0.39	0.43	0.34
Gemma	G3-4b	0.78	0.77	0.78	0.78	0.78	0.17	0.32	0.22	0.23	0.52
	G3-12b	0.78	0.79	0.79	0.79	0.78	0.34	0.50	0.38	0.48	0.50
Llama	L3.2-3b	0.79	0.78	0.78	0.79	0.79	0.32	0.39	0.33	0.39	0.32
	L3.1-8b	0.78	0.78	0.78	0.78	0.78	0.29	0.37	0.22	0.34	0.58

Table 7: **Image.** Test outcomes and mixed-panel agreement $\bar{R}_m$ for every image judge and prompting technique. See the caption of Table 6.

Family	Judge	Zero	Few	CoT	Ens	GEPA	Zero	Few	CoT	Ens	GEPA
		T2I-DPO ($\bar{R}_h = 0.12$)					**Vibe-Arena** ($\bar{R}_h = 0.12$)				
Gemini	Gem-P	**0.15**	**0.15**	0.14	**0.15**	✗	0.12	0.11	0.13	0.12	✗

Image (continued)

Family	Judge	Zero	Few	CoT	Ens	GEPA	Zero	Few	CoT	Ens	GEPA
		FER+ ($\bar{R}_h = 0.51$)					**Flickr8k-Expert** ($\bar{R}_h = 0.87$)				
	Gem-F	**0.15**	0.14	**0.15**	0.14	✗	0.13	0.11	0.11	0.12	✗
	Gem-L	0.13	0.12	0.13	0.12	✗	**0.14**	0.12	0.12	**0.14**	✗
GPT	GPT-T	0.12	0.13	0.14	0.13	✗	**0.14**	0.12	**0.14**	0.13	✗
	GPT-L	0.11	0.11	0.12	0.11	✗	0.11	0.08	0.11	0.11	✗
Qwen	Q2.5V-3b	−0.00	0.08	0.00	0.00	−0.00	−0.05	0.01	−0.05	−0.05	−0.06
	Q2.5V-7b	0.03	0.11	0.07	0.06	0.04	0.11	0.06	0.05	0.08	0.08
Gemma	G3-4b	0.06	0.08	0.01	0.05	−0.07	−0.06	0.07	−0.01	−0.03	−0.03
	G3-12b	0.11	0.08	0.04	0.10	0.11	0.06	0.10	−0.00	0.06	0.05
Llama	L3.2V-11b	−0.07	−0.06	0.02	−0.07	0.02	−0.04	−0.04	−0.01	−0.06	−0.06
		CIFAR-10H ($\bar{R}_h = 0.88$)									
Gemini	Gem-P	0.89	0.89	0.89	0.89	✗					
	Gem-F	**0.90**	**0.90**	**0.90**	**0.90**	✗					
	Gem-L	0.88	0.89	0.88	0.88	✗					
GPT	GPT-T	0.70	0.74	0.74	0.73	✗					
	GPT-L	0.67	0.74	0.65	0.70	✗					
Qwen	Q2.5V-3b	0.80	0.78	0.76	0.81	0.80					
	Q2.5V-7b	0.82	0.83	0.79	0.83	0.82					
Gemma	G3-4b	0.81	0.78	0.81	0.82	0.84					
	G3-12b	0.83	0.77	0.82	0.83	0.81					
Llama	L3.2V-11b	0.81	0.51	0.80	0.80	0.79					

Table 8: 🔊 **Audio.** Test outcomes and mixed-panel agreement $\bar{R}_m$ for every audio judge and prompting technique. See the caption of Table 6.

Family	Judge	Zero	Few	CoT	Ens	GEPA	Zero	Few	CoT	Ens	GEPA
		TTS-Preference ($\bar{R}_h = 0.01$)					**SOMOS** ($\bar{R}_h = 0.05$)				
Gemini	Gem-P	0.01	0.00	0.02	**0.03**	✗	−0.14	0.03	−0.12	−0.10	✗
	Gem-F	−0.02	−0.03	0.01	0.01	✗	**0.06**	0.04	0.04	0.03	✗
	Gem-L	−0.09	−0.03	−0.00	0.02	✗	−0.10	−0.04	0.05	−0.06	✗
GPT	GPT-A	0.02	−0.04	−0.03	−0.05	✗	0.04	0.04	0.05	0.04	✗
	GPT-A-M	−0.01	−0.00	−0.00	−0.00	✗	0.01	0.05	0.04	0.04	✗
Qwen	Q2.5-O-3b	−0.09	−0.07	0.01	0.01	−0.09	0.03	0.04	0.03	0.05	0.03
	Q2-A-7b	−0.12	−0.04	−0.12	−0.12	−0.12	−0.00	0.01	−0.00	0.02	−0.00

Table 9: **Video.** Test outcomes and mixed-panel agreement $\bar{R}_m$ for every video judge and prompting technique. See the caption of Table 6.

Family	Judge	Zero	Few	CoT	Ens
		T2V-Preference ($\bar{R}_h = 0.04$)			
Gemini	Gem-P	0.07	0.07	**0.08**	**0.08**
	Gem-F	**0.08**	**0.08**	**0.08**	**0.08**
	Gem-L	0.06	0.06	0.06	0.07
Qwen	Q2.5V-3b	0.03	−0.00	−0.01	0.03
	Q2.5V-7b	**0.08**	**0.08**	0.07	**0.08**

Table 10: **Outcomes of the three tests on datasets with shared annotators.** For each dataset and judge, the table reports the ALT-TEST winning rate ω at $\varepsilon = 0.1$, the ALTRA-TEST decision statistic q^*_α, the STT p_{TOST}, and the STT-Alt p_{lower} at $\mathrm{frac} = 0.5$, using each judge's best prompting technique. For each dataset, all tests are applied to the items that satisfy the requirements of every test. Green marks a passing cell. Because each test depends on its own hyperparameter, the comparison should not be read as strictly apples-to-apples.

	Gemini			GPT		Qwen		Gemma		Llama	
Text	**Gem-P**	**Gem-F**	**Gem-L**	**GPT-T**	**GPT-L**	**Q3-8b**	**Q3-14b**	**G3-4b**	**G3-12b**	**L3.2-3b**	**L3.1-8b**
DICES-350 ($\bar{R}_h = 0.20$)											
ALT-TEST ω	0.89	0.81	0.78	0.88	0.87	0.12	0.70	0.49	0.98	0.87	0.80
ALTRA-TEST q^*_α	+0.03	−0.07	−0.06	+0.01	+0.01	−0.09	−0.02	−0.06	+0.04	−0.03	−0.02
STT p_{TOST}	0.01	1.00	1.00	0.15	0.02	1.00	0.94	1.00	0.08	1.00	1.00
STT-Alt p_{lower}	< 0.01	1.00	1.00	0.15	0.02	1.00	0.94	1.00	< 0.01	1.00	1.00
EPIC ($\bar{R}_h = 0.20$)											
ALT-TEST ω	0.42	0.26	0.11	0.42	0.37	0.11	0.21	0.00	0.05	0.05	0.00
ALTRA-TEST q^*_α	+0.05	+0.01	+0.06	+0.05	+0.10	+0.08	+0.07	−0.09	+0.06	+0.01	−0.10
10k-Prompts ($\bar{R}_h = 0.37$)											
ALT-TEST ω	0.56	0.11	0.67	0.00	0.67	0.78	0.33	0.67	0.33	0.44	0.00
ALTRA-TEST q^*_α	+0.07	+0.05	+0.09	+0.01	+0.03	+0.05	+0.05	−0.01	+0.02	−0.04	−0.05
SummEval ($\bar{R}_h = 0.60$)											
ALT-TEST ω	1.00	0.67	0.33	0.33	0.33	1.00	1.00	0.00	1.00	0.00	0.67
ALTRA-TEST q^*_α	+0.05	+0.01	−0.02	−0.05	−0.07	−0.09	−0.04	−0.19	−0.04	−0.21	−0.10
ConvAbuse ($\bar{R}_h = 0.71$)											
ALT-TEST ω	0.75	0.25	0.50	0.25	0.50	0.50	0.12	0.00	0.00	0.00	0.00
ALTRA-TEST q^*_α	+0.01	−0.01	+0.00	−0.02	+0.02	−0.00	−0.04	−0.32	−0.10	−0.23	−0.14
CEBaB ($\bar{R}_h = 0.72$)											
ALT-TEST ω	0.60	0.90	1.00	0.50	0.80	0.50	0.60	0.70	0.90	0.50	0.50
ALTRA-TEST q^*_α	+0.07	+0.07	+0.07	+0.07	+0.07	+0.06	+0.06	+0.06	+0.07	+0.06	+0.06
Loneliness ($\bar{R}_h = 0.90$)											
ALT-TEST ω	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ALTRA-TEST q^*_α	−0.24	−0.28	−0.35	−0.35	−0.34	−0.49	−0.49	−0.43	−0.45	−0.56	−0.37
Image	**Gem-P**	**Gem-F**	**Gem-L**	**GPT-T**	**GPT-L**	**Q2.5V-3b**	**Q2.5V-7b**	**G3-4b**	**G3-12b**	**L3.2V-11b**	
Flickr8k-Expert ($\bar{R}_h = 0.87$)											
ALT-TEST ω	0.00	0.33	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	
ALTRA-TEST q^*_α	−0.14	−0.01	−0.04	−0.01	−0.04	−0.15	−0.09	−0.06	−0.05	−0.09	
Audio	**Gem-P**	**Gem-F**	**Gem-L**	**GPT-A**	**GPT-A-M**	**Q2.5-O-3b**	**Q2-A-7b**				
SOMOS ($\bar{R}_h = 0.05$)											
ALT-TEST ω	0.69	0.58	0.12	1.00	0.85	0.92	0.96				
ALTRA-TEST q^*_α	+0.01	+0.04	+0.04	+0.03	+0.04	+0.03	+0.01				

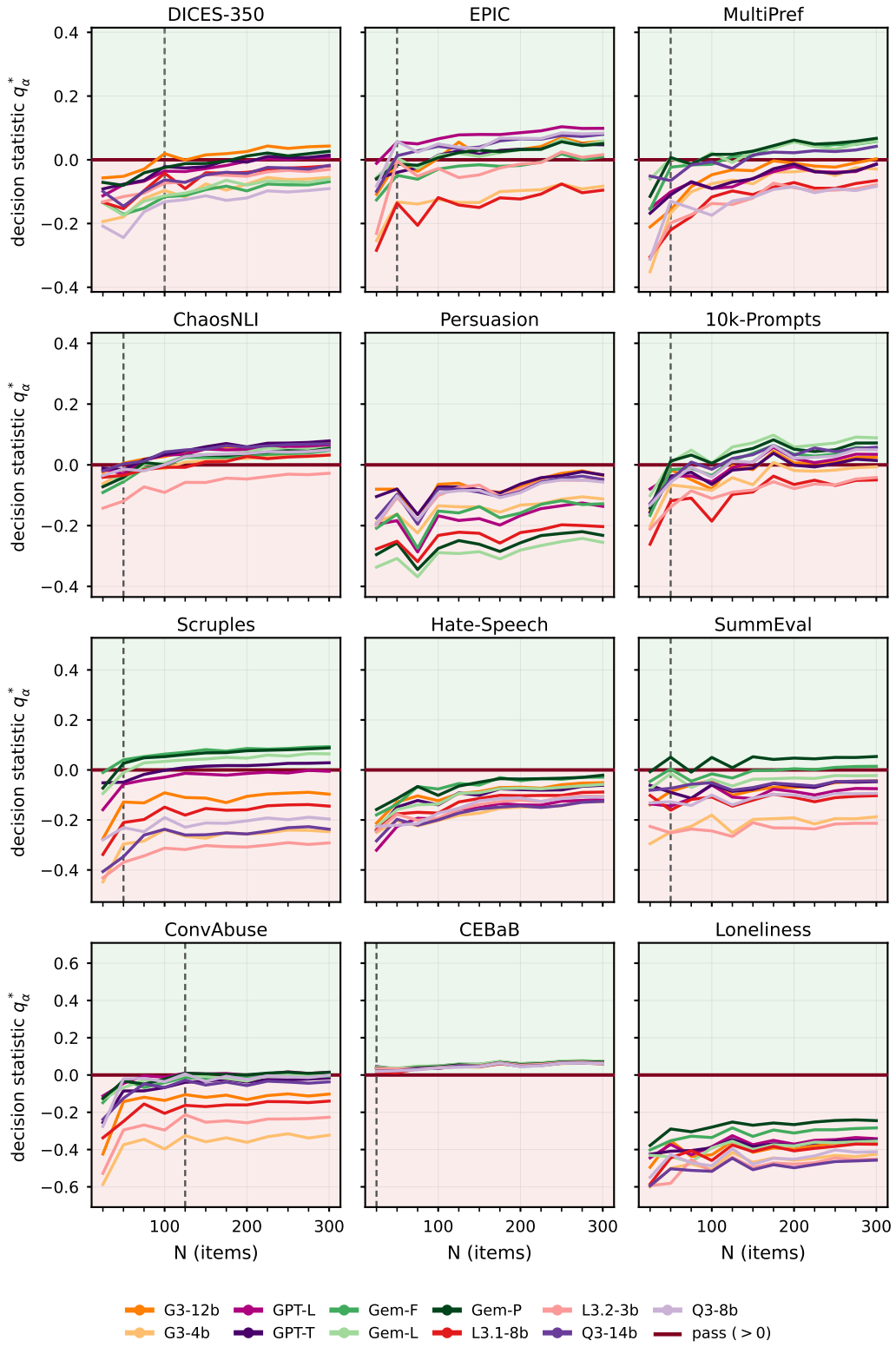


Figure 11: Varying the sample size N for real-world text datasets. The Y-axis of each plot presents the decision statistic q^*_α against the number of evaluated items N, with one curve per LLM, using each judge's best prompting technique. When $q^*_\alpha > 0$, the LLM passes the test. The dashed line marks the smallest N at which any judge first passes the test.

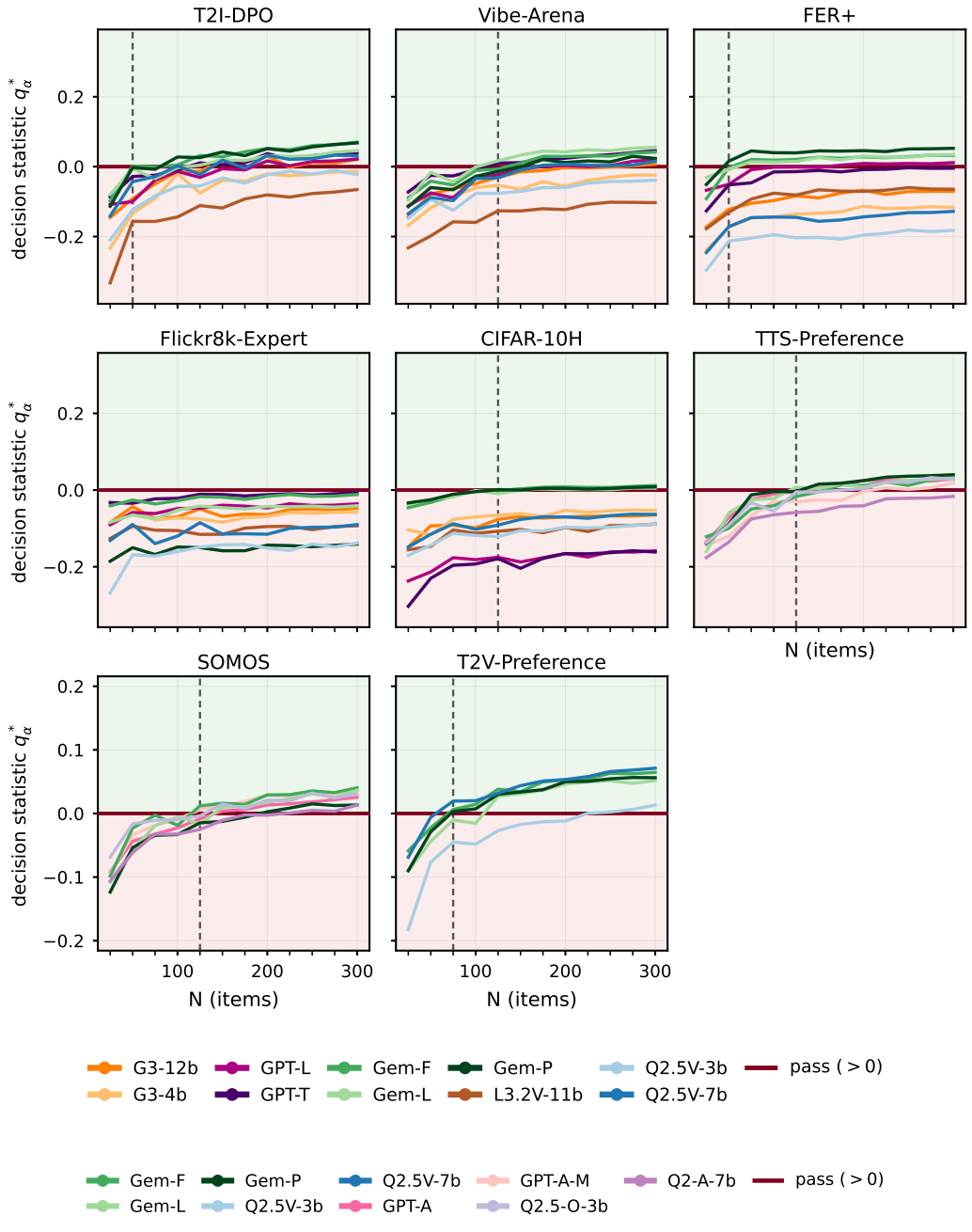


Figure 12: Varying the sample size N for real-world image, audio, and video datasets. See the caption of Figure 11.

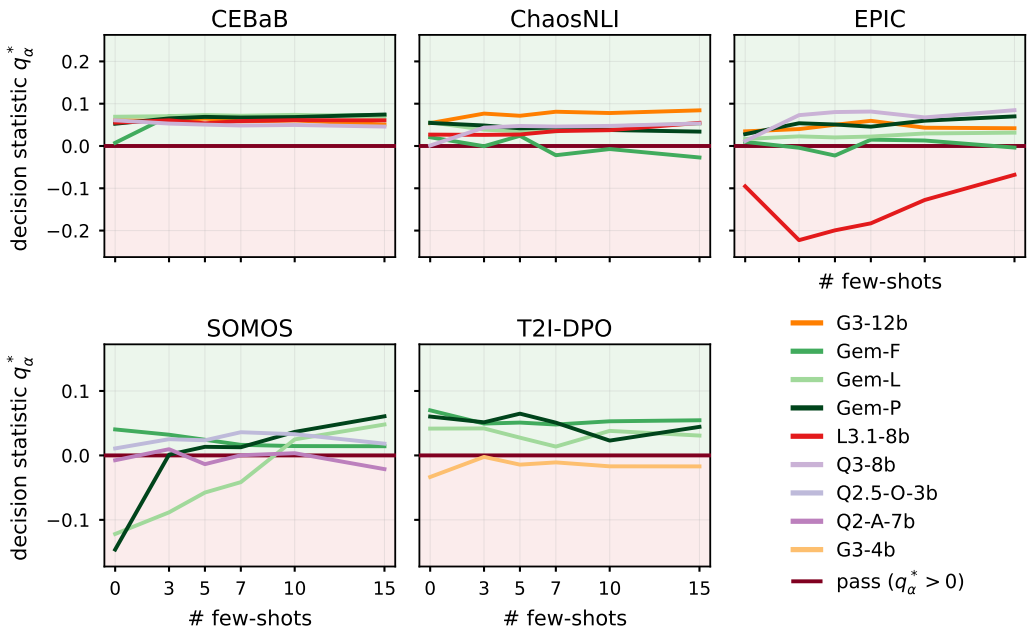


Figure 13: Evaluating the impact of the number of random few-shot examples (X-axis) on the decision statistic q^*_α. As shown, the effect is mixed and depends on both the model and the dataset.

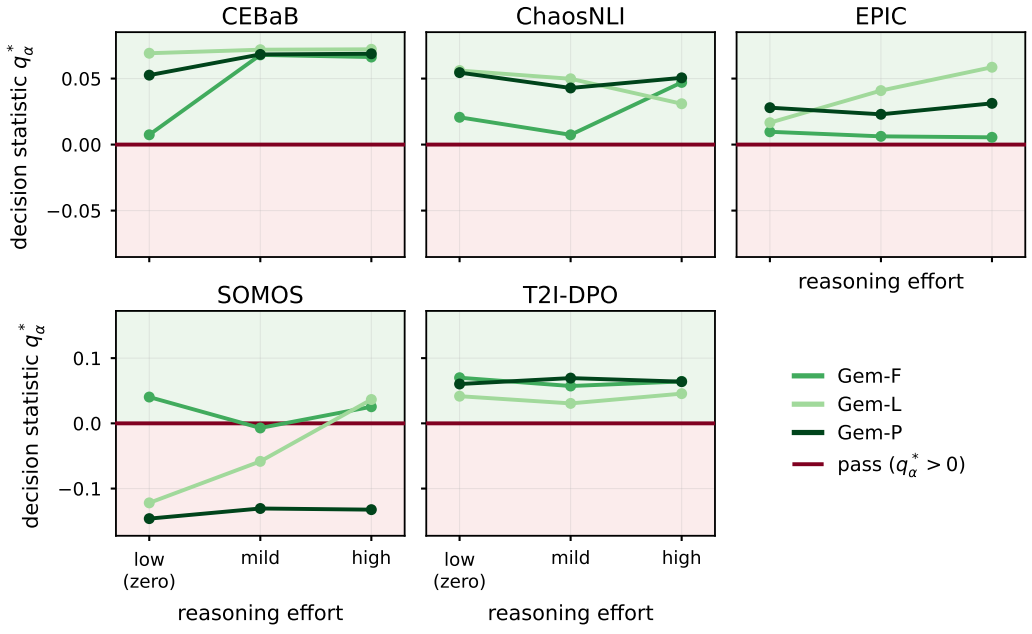


Figure 14: Evaluating the impact of reasoning effort in CoT prompting (X-axis) on the decision statistic q^*_α. Reasoning effort is controlled by the Gemini API thinking-effort parameter. As shown, the effect is mixed, but higher effort generally leads to better judges.