Rillor

METHODS NOTE

Measuring label quality with inter-annotator agreement

How percent agreement, Cohen's kappa, Fleiss' kappa and Krippendorff's alpha fit each annotation design, what inter-annotator agreement misses, and the label error rates published test sets carry.



8 November 2026 Labeling and annotation rillor.com/insights/measuring-label-quality-inter-annotator-agreement

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Inter-annotator agreement shows whether labeling is reproducible, and a label quality report also needs a measured error rate [1, 2].

- Inter-annotator agreement measures how far independent annotators give the same labels to the same items, beyond what chance would produce. High agreement shows a reproducible process and does not by itself show that the labels are right [1, 2].
- Percent agreement overstates reliability when one class dominates. If 95% of items belong to one class, two annotators labeling at those rates agree on 90.5% of items by chance, so an observed 90% is worse than chance [1].
- The coefficient should match the annotation design. A 2024 audit of 591 dataset papers recommends Krippendorff's alpha in almost all cases, Cohen's kappa for two annotators who label every item, and Fleiss' kappa only when annotators are randomly assigned to items [2].
- Spans, ratings and preference judgments need their own measures. Phrase-marking studies usually lack the count of negative cases that kappa needs [3], correlation does not show whether raters give similar scores [2], and random judges choosing between two responses agree half the time when ties are left out [4].
- Agreement is often reported without the details needed to judge it. In a 2024 audit of 591 dataset papers, 52% of those with human annotation reported agreement, 18% an error rate, and none a confidence interval for agreement [2].
- Published test sets carry measurable label errors. One 2021 study estimated at least 3.3% errors on average across 10 widely used test sets, and at least 6% in the ImageNet validation set [5].
- Disagreement can be part of the data. With 100 annotations per example, models scored near-perfect accuracy on items people agreed on and barely beat a random guess on items they did not [6].
- Model annotators need validation against human labels. Across 27 annotation tasks, one 2023 study found nine where a language model's labels had precision or recall below 0.5 [7].

Background

Labels are the answers a model learns from and is scored against, and checking them raises two questions. Reliability asks whether the labeling process is reproducible: in their survey of agreement coefficients, Artstein and Poesio call data reliable when coders can be shown to agree on the categories assigned to units to the extent the study's purpose requires [1]. Validity asks whether the labels are right. Agreement answers only the first, because two observers of the same event can share the same prejudice and both be wrong [1].

Agreement is still informative about the guidelines. When coders agree consistently, Artstein and Poesio infer that they have internalized a similar understanding of the annotation guidelines [1].

The coefficients come from content analysis, whose goals differ from those of annotation for computational linguistics [1]. Krippendorff writes that alpha applies wherever two or more methods of generating data are applied to the same set of objects [8].

Practice lags the methods. In a 2024 audit, Klie, Eckart de Castilho and Gurevych annotated 591 dataset papers from major language processing conferences for how each managed annotation quality [2]. They judged 30% of the papers' quality management "only subpar" and found common errors, especially in inter-annotator agreement and annotation error rates [2]. Of the papers with human annotators, 52% reported at least one agreement measure and 18% estimated an error rate [2]. To combine several labels for an item, 34% used majority voting, 14% manual curation and 1% probabilistic aggregation, and 46% did not say [2].

This note is for engineers, annotation leads and dataset buyers. It ties each coefficient to the design it fits, sets out what agreement cannot detect, reviews measured error rates and ends with what a quality note should report.

Method

What was examined. The definitions and uses of the agreement coefficients, the intraclass correlation coefficient (ICC) and pairwise F1, and the published evidence on label errors, error detection, annotator types and disagreement.

Sources. Thirty-three sources, each opened on 7 November 2026 and read for the passages cited, six of them through their abstract pages: 18 journal articles (two read in their arXiv versions), 9 peer-reviewed conference papers, 3 preprints, a technical paper by the author of Krippendorff's alpha and 2 software documentation pages.

How. Each coefficient is described from the paper that defines it, or as the survey, a statistics tutorial and software documentation state it [1, 9, 10, 11]. Figures are as each source prints them. Results for model annotators are kept to the tasks and dates each study tested, and the models are described as commercial model families.

Where sources disagree. The audit's text counts three datasets with probabilistic aggregation and its Table 1 counts 2 papers [2]. This note uses the table. McHugh attributes the scale from "slight" to "almost perfect" to Cohen [12], and the survey, the audit and Hallgren, followed here, to Landis and Koch (1977) [1, 2, 9]. The test-set study's abstract says label errors make up at least 6% of the ImageNet validation set, and its Table 1 gives 5.83%, 2,916 of 50,000 [5].

Findings

1. Percent agreement, kappa and alpha

Percent agreement. Observed agreement is the share of items two coders labeled the same. In the survey's example of two coders and 100 utterances, it is (20 + 50)/100 = 0.7 [1]. It is easy to read and biased. Quoting Scott (1955), the survey notes that percentage agreement favors dimensions with a small number of categories: coders choosing at random between two equally likely categories agree on half the items, and with three on a third [1]. Skewed classes raise chance agreement further. If 95% of utterances belong to one class, coders labeling at those rates agree on 90.5% by chance, so "a seemingly high observed agreement of 90%" is worse than chance [1].

Chance correction. The chance-corrected coefficients share one form: observed agreement minus expected agreement, divided by one minus expected agreement, where 0 means agreement at chance level and 1 perfect agreement [1]. scikit-learn's documentation states Cohen's kappa as $\kappa = (p_o - p_e) / (1 - p_e)$, with p_e estimated from each annotator's own distribution over the labels [10]. S, Scott's pi and Cohen's kappa differ only in expected agreement. S assumes all categories are equally likely, pi uses one distribution pooled across coders, and kappa uses a separate distribution for each coder [1].

The choice of coefficient moves the number. On one table of two coders, three categories and 100 items with observed agreement 0.88, the survey computes S at 0.82, pi at about 0.7995, kappa at about 0.8013, weighted alpha at about 0.8156 and weighted kappa at about 0.8163 [1]. All five sit near 0.8, the value the survey suggests where a threshold is needed [1]. Pi falls just under that line and kappa just over it [1]. In McHugh's worked example of 222 items, percent agreement is 0.94 and kappa 0.85, with a 95% interval of 0.78 to 0.92 [12].

Kappa's paradoxes. For a fixed observed agreement, a larger difference between the two coders' label distributions gives a lower expected agreement and so a higher kappa, which Di Eugenio and Glass called a paradox [1]. The survey prefers multi-coder pi and alpha because they abstract away from the bias of individual coders, and notes that the differences shrink as annotators are added [1].

The second paradox is prevalence: chance-corrected agreement can be low even when observed agreement is high [1]. Feinstein and Cicchetti showed that a substantial imbalance in a table's marginal totals can drastically lower kappa computed from a high observed agreement [13], and Hallgren notes that the prevalence problem typically makes kappa estimates unrepresentatively low [9]. The survey's answer is that the test of reliability in such cases is the ability to agree on the rare categories [1]. Gwet proposed a coefficient called AC1 as an alternative and more stable measure for these paradoxes [14].

Krippendorff's alpha. Alpha is one minus the ratio of observed to expected disagreement. A value of 1 is perfect reliability and 0 indicates the absence of reliability [8]. It applies to any number of observers, any level of measurement (nominal, ordinal, interval, ratio and more), incomplete or missing data and small or large samples, and it evaluates reliability one variable at a time [8]. When all disagreements count equally, alpha is

nearly identical to multi-coder pi, with a correction for small samples [1]. The metric has to match the label type. In Krippendorff's example of 4 observers and 12 units with 7 of 48 possible values missing, the same data give alpha of 0.743 as nominal labels, 0.815 as ordinal, 0.849 as interval and 0.797 as ratio [8].

Fleiss' kappa and weighted kappa. Statistics for three or more coders on nominal data are typically extensions of Scott's pi, such as Fleiss' 1971 statistic [9]. The statsmodels implementation uses the sample margin to define chance [11]. Weighted kappa penalizes disagreements by their size, for ordered categories such as high, medium and low [9], with linear or quadratic weights in scikit-learn [10]. The survey warns that choosing the weights makes it harder to decide when a value is sufficient [1].

2. Matching the coefficient to the annotation design

Design questions come first. Hallgren asks whether all items were rated by several coders or only a subset, and whether the same coders rated every item, a fully crossed design, or different coders rated different items [9]. Fully crossed designs allow systematic bias between coders to be assessed and controlled for, and Cohen's original kappa is only suitable for fully crossed designs with exactly two coders [9].

The audit's recommendation. Klie and colleagues recommend Krippendorff's alpha because it "can be used in almost all circumstances", including sequence tagging in its unitized form, continuous judgments and varying numbers of annotations per item [2]. When the same number of annotators label each item, Cohen's kappa for two annotators or Fleiss' kappa for more can be added, the latter only if annotators are randomly assigned to items [2]. Percent agreement should rarely be used and never as the only measure [2]. The survey also recommends alpha and calls it indispensable when labels are not equally distinct from one another, as in hierarchical tag sets, while noting that the choice of weights affects the result [1].

Annotation design	Coefficient the sources fit to it	Basis
Two annotators label every item, unordered labels	Cohen's kappa, or alpha	[2, 9]
The same number of annotators per item, randomly assigned	Fleiss' kappa, or alpha	[2, 11]
Varying annotators per item or missing labels	Krippendorff's alpha	[2, 8]
Ordered labels such as high, medium and low	Weighted kappa or ordinal alpha	[8, 9, 10]
Scores on a numeric scale	ICC or interval alpha	[2, 8, 15]
Entities and spans in text	Pairwise F1, or unitized alpha	[2, 3, 16]
Choices between two responses	Raw agreement read against its chance level	[4]

Spans and entities. Studies that mark phrases usually lack a well-defined number of negative cases, which prevents the use of kappa, and Hripcsak and Rothschild show that the pairwise F-measure equals positive specific agreement, with kappa approaching it as negatives grow numerous [3]. A medical corpus study found that token-level kappa does not properly reflect the annotation task and is computed on very imbalanced data [17]. In an archaeology entity dataset, one 100-sentence test document gave Cohen's kappa of 0.82 on all tokens, 0.67 on annotated tokens only and an F1 of 0.95 [16]. Its authors report that pairwise F1, leaving out tokens outside any entity, is usually seen as a better measure there [16]. The audit notes that F1 is not chance-corrected and compares only pairs, and it found one paper that used unitized alpha for sequence labeling [2]. Unitized alpha assumes that one coder's units do not overlap [1].

Ratings on a scale. Correlation measures whether two raters' scores move together. It does not measure whether the scores are similar, and two raters with opposite biases can show high correlation and low agreement [2]. Seven papers in the audit (2%) still used correlation as agreement [2]. The audit points to the ICC instead [2], and interval alpha fits the same data [8]. Shrout and Fleiss describe six forms of the ICC for targets rated by several judges [15]. Koo and Li read the 95% interval of an ICC as poor below 0.5, moderate from 0.5 to 0.75, good from 0.75 to 0.9 and excellent above 0.9, with a rule of thumb of at least 30 heterogeneous samples and at least 3 raters [18].

Preference judgments. Choices between two model responses have their own chance level. In MT-Bench, random judges agree 33% of the time with ties counted and 50% without ties, and expert human judges agreed with each other 63% with ties and 81% without ties on first turns [4]. In an instruction-following study, training labelers agreed with each other 72.6 ± 1.5% of the time and held-out labelers 77.3 ± 1.3% [19]. These are raw agreement rates on comparisons, to be read against the chance level of the comparison format [4].

3. What agreement does not tell you

Reliability is necessary and is not sufficient. The audit states that high agreement "does not automatically guarantee high-quality labels" and only demonstrates a reliable annotation process [2]. Agreement also does not show whether the scheme and guidelines capture the phenomena a project wants [2]. Passonneau and Carpenter call standard agreement measures neither necessary nor sufficient to ensure a high-quality corpus, and their annotation model gives a certainty measure for each gold label instead [20]. In their word-sense case, crowdsourced labels with the Dawid and Skene model cost less than half as much as the conventional approach, and most items had high certainty even where trained annotators' agreement was low [20].

Agreement does not predict learning. The survey summarizes Reidsma and Carletta's finding that agreement coefficients are poor predictors of machine-learning success [1]. Reidsma and Carletta show that low reliability is tolerable as long as disagreement looks like random noise, while patterned disagreement can make performance figures look better than they are, so even a measure of 0.8 does not guarantee that good-looking performance is real [21].

Thresholds are conventions. The ranges of Landis and Koch (1977) read 0.01 to 0.20 as slight agreement, 0.21 to 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial and 0.81 to 1.00 almost perfect [2, 9]. McHugh proposes a stricter table, with 0.80 to 0.90 read as strong [12]. The survey quotes Carletta that content analysts treat values above 0.8 as good reliability, with 0.67 to 0.8 allowing tentative conclusions, and Krippendorff calling a cutoff of 0.800 "a pretty low standard" [1]. The audit reports that Krippendorff later called 0.8 the absolute lower limit that should better be 0.9, and that alpha below 0.667 is said to indicate unreliable labels [2]. The survey advises 0.8 if a threshold has to be set and doubts that one cutoff suits all purposes [1]. The audit calls such boundaries arbitrary, notes that Landis and Koch considered only binary classification, and recommends against using thresholds from the literature [2].

Misuse in practice. In the audit's corpus, the agreement value was often stated without any interpretation or comment (52%) [2]. Authors used freeform labels in 27% of cases, and values below 0.3 were described as reasonable [2]. Cohen's kappa was reported where more than two annotations per item were collected, and kappa or alpha was sometimes given in percent [2]. About 11% of the papers that used any measure reported percent agreement alone, and the authors found no confidence interval for any agreement value [2].

Sample size. Where agreement was computed on a subset of the data, the median subset was 200 items, and 51% of those values came from 200 items or fewer and 28% from 100 or fewer [2]. For an expected agreement of 0.8 with a precision of ±0.05 at 95% confidence, the audit puts the minimum at about 500 annotated items, and it recommends a confidence interval, for example by bootstrapping [2].

Independence. Krippendorff requires coders who work independently, and discussion or consensus by majority vote makes the data unusable for measuring reproducibility [1]. He asks for three or more coders working independently from a written scheme fixed in advance, so expanding the instructions during labeling breaks the condition [1]. The reliability sample should let each category yield at least five chance agreements and contain examples of every category [1].

Self-consistency is a different measure. In one 2023 study, intercoder agreement for a commercial chat model compared two runs of the same model [22]. Average agreement was about 56% for workers on a crowdsourcing platform, 79% for trained annotators, 91% for the model at temperature 1 and 97% at temperature 0.2 [22]. The model's figures measure run-to-run consistency, and accuracy in that study meant agreement with the trained annotators [22].

4. Label errors in published test sets

Northcutt, Athalye and Mueller estimated an average of at least 3.3% label errors across 10 widely used test sets, and at least 6% in the ImageNet validation set [5]. Error shares ranged from 0.15% in MNIST to 10.12% in QuickDraw [5].

Test set	Estimated label errors	Test set	Estimated label errors
MNIST	0.15%	QuickDraw	10.12%

Test set	Estimated label errors	Test set	Estimated label errors
CIFAR-10	0.54%	20news	1.09%
CIFAR-100	5.85%	IMDB	2.90%
Caltech-256	1.54%	Amazon Reviews	3.90%
ImageNet	5.83%	AudioSet	1.35%

Source: [5], Table 1.

How the errors were confirmed. Each flagged example was shown independently to five workers, and it counted as an error when fewer than three agreed with its given label [5]. On average 51% of the flagged candidates were confirmed [5]. Of ImageNet's 2,916 confirmed errors, 1,428 were correctable, 597 multi-label, 293 in a "neither" category and 598 in "non-agreement" [5]. A label error in this sense includes items with no single right label [5]. The authors add that datasets with more carefully designed labeling methods and more careful human curation generally had less error [5].

Effect on model choice. The authors find that lower-capacity models may be more useful than higher-capacity ones when a large share of test labels is wrong [5]. On corrected ImageNet labels, ResNet-18 outperforms ResNet-50 if the share of originally mislabeled test examples rises by just 6% [5]. Beyer and colleagues relabeled the ImageNet validation set with a more robust procedure and found recent classifiers' gains substantially smaller than on the original labels [23].

What dataset authors report. Only 18% of the audit's papers with human annotation reported an error rate, with an average of 8.27% and a median of 6.00% [2]. Of 80 error rates, 64 came from a subset with a median of 200 items, and no work chose its sample size on a statistical footing [2]. A small sample can mislead. The authors of one relation-classification dataset inspected 300 instances and estimated an error rate of about 6.7%, and a later full inspection found 23.9% [2]. The audit recommends inspecting at least 500 instances or the whole dataset, whichever is smaller, and stating a confidence interval [2]. For a true error rate of 5%, a sample of 456 gives a 95% interval of about ±0.02 [2].

5. Finding label errors at scale

Confident learning. Northcutt, Jiang and Chuang characterize and identify label errors by pruning noisy data, counting with probabilistic thresholds and ranking examples [24]. The method estimates the joint distribution of given and true labels under a class-conditional noise assumption, and it is not tied to a data type or model [24]. On ImageNet it estimated that 645 "missile" images are mislabeled as their parent class

"projectile", a case of overlapping classes [24]. The method is available as the open-source cleanlab release [24]. A flag is a candidate for review: in the test-set study, about half of the flagged items were confirmed by people [5].

A benchmark of detectors. Klie, Webber and Gurevych reimplemented 18 methods for detecting annotation errors and ran them on 9 English datasets for text classification and token and span labeling [25]. Four came out ahead overall: Borda Count, Diverse Ensemble, Label Aggregation and Retag [25]. Inconsistencies were harder to detect for most methods, especially model-based ones [25].

Annotator models. MACE learns without supervision which annotators are trustworthy and predicts the correct underlying labels [26]. Its authors built it for annotators who choose bad labels to maximize their pay, a strategy that defeats majority voting when several annotators follow it [26]. The Dawid and Skene model gives each annotator a confusion matrix over the label classes, and the audit notes that majority voting is often a strong baseline [2].

6. Crowd, expert and model annotators

Crowd and experts. In a 2008 study, an average of 4 non-expert labels per example matched the agreement of a single expert for affect recognition, at 2008 rates of at least 875 expert-equivalent labels per US dollar [27]. Voting over 10 annotators reached 0.94 agreement with the gold labels on temporal ordering, and on a word-sense task one disagreement with the gold standard turned out to be an error in the gold annotation [27]. The survey warns that agreement among long-term collaborators can come from their knowledge of the research purpose instead of from the written instructions [1]. It still accepts expert coders: agreement studies assure that results replicate when annotators come from the same population as the original ones, alongside an independent evaluation of the task the corpus serves [1].

Model annotators. In one 2023 study on four samples of tweets and news articles (6,183 documents), a commercial chat model's zero-shot accuracy exceeded crowd workers' on most tasks, by about 25 percentage points on average, at less than $0.003 per annotation, about thirty times cheaper than the crowdsourcing platform [22]. The tasks were relevance, stance, topic and frame detection [22]. A second 2023 study tested 27 annotation tasks across 11 datasets and slightly over 200,000 text samples [7]. It found a median accuracy of 0.850 and a median F1 of 0.707, with F1 from 0.059 to 0.969, and nine of the 27 tasks had precision or recall below 0.5 [7]. Its authors conclude that any automated annotation by a language model must be validated against labels produced by people [7].

Calderon, Reichart and Dror propose a test before replacing people: compare the model with a small group of at least three human annotators on 50 to 100 examples [28]. Their results show that language models can sometimes replace humans [28]. As judges of chat responses, a commercial model's judgments agreed with expert humans on 85% of non-tied votes, against 81% between the humans themselves, and the same paper studies position, verbosity and self-enhancement biases in model judges [4].

7. When disagreement is data

Plank argues that human label variation affects data, modeling and evaluation [29]. Conventional practice assumes a ground truth and neglects genuine variation from disagreement, subjectivity or several plausible answers, and the paper provides a repository of public datasets with unaggregated labels [29].

Pavlick and Kwiatkowski collected 50 judgments per sentence pair and found that disagreements persist as more ratings are collected and as the context given to raters varies [30].

ChaosNLI collected 100 annotations per example for 3,113 examples in SNLI and MNLI and 1,532 in αNLI, 464,500 in all [6]. The old majority label differed from the new 100-vote majority for 10.64% of the αNLI examples and 24.97% and 31.77% of the SNLI and MNLI subsets, which were chosen for low agreement [6]. Models reached near-perfect accuracy on items with high human agreement and barely beat a random guess on items with low agreement [6].

Uma and colleagues survey training from multiple judgments that may disagree, and find that the relative performance of the training methods is critically affected by the form of evaluation chosen [31]. The audit adds that low agreement does not automatically mean low-quality labels, since some tasks are inherently subjective, and that majority voting might drown out valid minority voices [2].

Implications

Drawing on the reporting lists of the survey and the audit [1, 2], a label quality note for a dataset can state:

1. **The design.** How many annotators labeled each item, how they were assigned, whether they worked independently and whether they relied only on a written guideline fixed in advance [1, 2, 9].
2. **The coefficient and why.** Alpha with its metric, kappa where the design fits, pairwise F1 with its matching rule for spans, or the ICC form for scales [2, 3, 8, 15].
3. **The sample.** Its size, how it was drawn and a confidence interval on the agreement value. About 500 items estimate an agreement near 0.8 within ±0.05 at 95% confidence [2].
4. **The table behind the number.** A confusion matrix or agreement table, so that an overall figure does not hide disagreement on less common categories [1], and agreement per label and per annotator, which may identify poorly performing annotators or labels that are hard to decide [2].
5. **The error rate.** Errors found after adjudication on a stated sample of at least 500 items or the whole dataset, whichever is smaller, with an interval [2].
6. **The aggregation.** Majority vote, curation or a probabilistic model, and whether the unaggregated labels are kept where disagreement may be real [2, 6, 29].
7. **Model-assisted labels.** Where a model proposed or set labels, its validation against human labels on the same task [7, 28].

Datasheets for datasets ask related questions: whether each instance has a label, whether any labeling was done, whether the raw data was saved, and whether the dataset is to be updated, for example to correct labeling errors [32]. Data statements also ask for the demographics of annotators [33].

For readers of such notes, a single coefficient near a published threshold says little without the design, the sample and the table behind it [1, 2]. Agreement cannot show that the labels are right [1, 2], so the note pairs it with an error rate measured on a reviewed sample [2].

Limits of this analysis

- Most of the evidence comes from language datasets in English: the audit covers papers from language processing conferences and the detector benchmark uses 9 English datasets [2, 25]. The test-set study covers image, text and audio [5].
- The test-set error rates are lower bounds, from one detection method plus confirmation by crowd workers [5].
- The model-annotator results come from particular model versions tested in 2023 and 2025, and hold for their tasks and dates [4, 7, 22, 28].
- Prices are dated: 2008 crowdsourcing rates in one study and 2023 rates in another [22, 27].
- Threshold tables are conventions, as their own sources say [1, 2].

Sources

1. Survey Article: Inter-Coder Agreement for Computational Linguistics. Ron Artstein and Massimo Poesio, Computational Linguistics 34(4), pp. 555-596, MIT Press, December 2008, viewed 7 November 2026. https://aclanthology.org/J08-4004/

2. Analyzing Dataset Annotation Quality Management in the Wild. Jan-Christoph Klie, Richard Eckart de Castilho and Iryna Gurevych, Computational Linguistics 50(3), pp. 817-866, September 2024, viewed 7 November 2026. https://aclanthology.org/2024.cl-3.1/

3. Agreement, the F-Measure, and Reliability in Information Retrieval. George Hripcsak and Adam S. Rothschild, Journal of the American Medical Informatics Association 12(3), pp. 296-298, May-June 2005, viewed 7 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC1090460/

4. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Lianmin Zheng and 12 others, NeurIPS 2023 Datasets and Benchmarks Track, arXiv 2306.05685, version 4, 24 December 2023, viewed 7 November 2026. https://arxiv.org/abs/2306.05685

5. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. Curtis G. Northcutt, Anish Athalye and Jonas Mueller, NeurIPS 2021 Datasets and Benchmarks Track, arXiv 2103.14749, version 4, 7 November 2021, viewed 7 November 2026. https://arxiv.org/abs/2103.14749

6. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? Yixin Nie, Xiang Zhou and Mohit Bansal, Proceedings of EMNLP 2020, pp. 9131-9143, Association for Computational Linguistics, November 2020, viewed 7 November 2026. https://aclanthology.org/2020.emnlp-main.734/

7. Automated Annotation with Generative AI Requires Validation. Nicholas Pangakis, Samuel Wolken and Neil Fasching, preprint, arXiv 2306.00176, version 1, 31 May 2023, viewed 7 November 2026. https://arxiv.org/abs/2306.00176

8. Computing Krippendorff's Alpha-Reliability. Klaus Krippendorff, Annenberg School for Communication, University of Pennsylvania, 25 January 2011, viewed 7 November 2026. https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf

9. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Kevin A. Hallgren, Tutorials in Quantitative Methods for Psychology 8(1), pp. 23-34, 2012, author manuscript in PubMed Central, viewed 7 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC3402032/

10. sklearn.metrics.cohen_kappa_score. scikit-learn developers, scikit-learn 1.9.1 documentation, undated page, viewed 7 November 2026. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

11. statsmodels.stats.inter_rater.fleiss_kappa. statsmodels developers, statsmodels 0.15.0 documentation, undated page, viewed 7 November 2026. https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

12. Interrater reliability: the kappa statistic. Mary L. McHugh, Biochemia Medica 22(3), pp. 276-282, 15 October 2012, viewed 7 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/

13. High agreement but low kappa: I. The problems of two paradoxes. Alvan R. Feinstein and Domenic V. Cicchetti, Journal of Clinical Epidemiology 43(6), pp. 543-549, 1990, PubMed abstract 2348207, viewed 7 November 2026. https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=2348207&rettype=abstract&retmode=text

14. Computing inter-rater reliability and its variance in the presence of high agreement. Kilem L. Gwet, British Journal of Mathematical and Statistical Psychology 61(1), pp. 29-48, May 2008, PubMed abstract 18482474, viewed 7 November 2026. https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=18482474&rettype=abstract&retmode=text

15. Intraclass correlations: uses in assessing rater reliability. Patrick E. Shrout and Joseph L. Fleiss, Psychological Bulletin 86(2), pp. 420-428, March 1979, PubMed abstract 18839484, viewed 7 November 2026. https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=18839484&rettype=abstract&retmode=text

16. Creating a Dataset for Named Entity Recognition in the Archaeology Domain. Alex Brandsen, Suzan Verberne, Milco Wansleeben and Karsten Lambers, Proceedings of LREC 2020, pp. 4573-4577, European Language Resources Association, May 2020, viewed 7 November 2026. https://aclanthology.org/2020.lrec-1.562/

17. Building Gold Standard Corpora for Medical Natural Language Processing Tasks. Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser and others, AMIA Annual Symposium Proceedings 2012, pp. 144-153, 3 November 2012, viewed 7 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC3540456/

18. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Terry K. Koo and Mae Y. Li, Journal of Chiropractic Medicine 15(2), pp. 155-163, 2016, viewed 7 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC4913118/

19. Training language models to follow instructions with human feedback. Long Ouyang and 19 others, preprint, arXiv 2203.02155, version 1, 4 March 2022, viewed 7 November 2026. https://arxiv.org/abs/2203.02155

20. The Benefits of a Model of Annotation. Rebecca J. Passonneau and Bob Carpenter, Transactions of the Association for Computational Linguistics 2, pp. 311-326, October 2014, viewed 7 November 2026. https://aclanthology.org/Q14-1025/

21. Squibs: Reliability Measurement without Limits. Dennis Reidsma and Jean Carletta, Computational Linguistics 34(3), pp. 319-326, MIT Press, 2008, viewed 7 November 2026. https://aclanthology.org/J08-3001/

22. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Fabrizio Gilardi, Meysam Alizadeh and Maël Kubli, Proceedings of the National Academy of Sciences 120(30), e2305016120, 2023, arXiv 2303.15056, version 2, 19 July 2023, viewed 7 November 2026. https://arxiv.org/abs/2303.15056

23. Are we done with ImageNet? Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai and Aäron van den Oord, preprint, arXiv 2006.07159, version 1, 12 June 2020, viewed 7 November 2026. https://arxiv.org/abs/2006.07159

24. Confident Learning: Estimating Uncertainty in Dataset Labels. Curtis Northcutt, Lu Jiang and Isaac Chuang, Journal of Artificial Intelligence Research 70, pp. 1373-1411, 14 April 2021, viewed 7 November 2026. https://www.jair.org/index.php/jair/article/view/12125

25. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. Jan-Christoph Klie, Bonnie Webber and Iryna Gurevych, Computational Linguistics 49(1), pp. 157-198, March 2023, viewed 7 November 2026.

https://aclanthology.org/2023.cl-1.4/

26. Learning Whom to Trust with MACE. Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani and Eduard Hovy, Proceedings of NAACL-HLT 2013, pp. 1120-1130, Association for Computational Linguistics, June 2013, viewed 7 November 2026. https://aclanthology.org/N13-1132/

27. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng, Proceedings of EMNLP 2008, pp. 254-263, Association for Computational Linguistics, October 2008, viewed 7 November 2026. https://aclanthology.org/D08-1027/

28. The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. Nitay Calderon, Roi Reichart and Rotem Dror, Proceedings of ACL 2025 (Volume 1: Long Papers), pp. 16051-16081, Association for Computational Linguistics, July 2025, viewed 7 November 2026. https://aclanthology.org/2025.acl-long.782/

29. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Barbara Plank, Proceedings of EMNLP 2022, pp. 10671-10682, Association for Computational Linguistics, December 2022, viewed 7 November 2026. https://aclanthology.org/2022.emnlp-main.731/

30. Inherent Disagreements in Human Textual Inferences. Ellie Pavlick and Tom Kwiatkowski, Transactions of the Association for Computational Linguistics 7, pp. 677-694, 2019, viewed 7 November 2026. https://aclanthology.org/Q19-1043/

31. Learning from Disagreement: A Survey. Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank and Massimo Poesio, Journal of Artificial Intelligence Research 72, pp. 1385-1470, 27 December 2021, viewed 7 November 2026. https://www.jair.org/index.php/jair/article/view/12752

32. Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford, Communications of the ACM, December 2021, arXiv 1803.09010, version 8, 1 December 2021, viewed 7 November 2026. https://arxiv.org/abs/1803.09010

33. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Emily M. Bender and Batya Friedman, Transactions of the Association for Computational Linguistics 6, pp. 587-604, 2018, viewed 7 November 2026. https://aclanthology.org/Q18-1041/

About this report

Published by Rillor on 8 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Measuring label quality with inter-annotator agreement. 8 November 2026. https://rillor.com/insights/measuring-label-quality-inter-annotator-agreement

Work with Rillor

Send a sample of the records, the labels your schema sets and the accuracy you need. Contact us at rillor.com/contact?type=labeling or contact@rillor.com. More on labeling and annotation: rillor.com/data#labeling