Counting on Consensus: Selecting the Right Inter-annotator Agreement Metric for NLP Annotation and Evaluation

Joseph James
Department of Computer Science, The University of Sheffield
Sheffield, United Kingdom
jhfjames1@sheffield.ac.uk

Abstract

Human annotation remains the foundation of reliable and interpretable data in Natural Language Processing (NLP). As annotation and evaluation tasks continue to expand, from categorical labelling to segmentation, subjective judgment, and continuous rating, measuring agreement between annotators has become increasingly more complex. This paper outlines how inter-annotator agreement (IAA) has been conceptualised and applied across NLP and related disciplines, describing the assumptions and limitations of common approaches. We organise agreement measures by task type and discuss how factors such as label imbalance and missing data influence reliability estimates. In addition, we highlight best practices for clear and transparent reporting, including the use of confidence intervals and the analysis of disagreement patterns. The paper aims to serve as a guide for selecting and interpreting agreement measures, promoting more consistent and reproducible human annotation and evaluation in NLP.

Keywords: inter-annotator agreement, annotation reliability, reproducibility, human evaluation

1. Introduction

Creating high-quality annotated data and ensuring reliable human evaluation are central to NLP, and the consistency of human judgments determines the validity of datasets and evaluations. IAA measures this consistency by quantifying the extent to which multiple annotators or evaluators apply the same labels to the same items. High IAA suggests clear guidelines and a reproducible results from texts or model outputs to human judgments, whereas low IAA can indicate under-specified instructions, insufficient training, or subjectivity in the task being evaluated (Hallgren, 2012). However, relying on raw agreement alone can overestimate reliability, motivating the use of chance-corrected and task-appropriate statistics (Hayes and Krippendorff, 2007; Kohen, 1960). The diversity of NLP tasks, from simple categorical labelling to span-based extraction, segmentation, and pairwise preferences, makes selecting the right agreement metric challenging. Recent work has highlighted how metric choice and distance functions affect interpretability and robustness for complex annotation tasks (Braylan et al., 2022), and evaluation studies document substantial variability in human criteria and correlations with automatic metrics (Chhun et al., 2022). This highlights the need to match the agreement coefficient to the annotation design and to clarify what construct the metric is intended to capture (Chhun et al., 2022).

Another concern is the reporting practices. Point estimates without uncertainty overstate precision and reduce comparability. Methodological guidance recommends reporting confidence intervals, describing the rater design, and accounting for missing data and label prevalence, all of which affect interpretation of the results (Hallgren, 2012; Hayes and Krippendorff, 2007). These considerations extend to human evaluation of model outputs, where pairwise preference studies reveal inconsistency under common setups and call for better design and meta-evaluation (Ghosh et al., 2024). Similar work shows that evaluator biases and inconsistencies can influence conclusions about system quality, meaning careful design is required to obtain reliable human judgments (Bavaresco et al., 2025; Liu et al., 2024; Balloccu et al., 2024; Tam et al., 2024). Finally, disagreement is not merely “noise”, modelling annotator or evaluator identities and perspectives can improve downstream learning and preserve legitimate variation in judgments, rather than collapsing it during aggregation (Deng et al., 2023).

This paper offers an overview of how IAA can be measured in NLP. It reviews commonly used approaches across different settings, describing how their underlying assumptions affect interpretation and comparability. In addition, it outlines methodological considerations for reporting, including how to account for data imbalance, missing annotations, and uncertainty in agreement estimates. The aim is to help researchers select measures that align with their task and interpret them appropriately in context. By treating agreement not as an afterthought but as a key component, this work seeks to support more transparent and reproducible practices in NLP.

2. Categorical data

For tasks where each item is assigned a category, agreement can be measured in two main ways:

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4434–4446
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

by directly calculating the proportion of matching labels between annotators, or by using chance-corrected coefficients that adjust for agreement expected by random chance.

2.1. Percentage Agreement

Percentage agreement (P_o) is the simplest and most direct measure of inter-annotator reliability. It represents the proportion of items for which annotators assign the same label and is defined as the number of agreed items divided by the total number of items. However, because it does not account for agreement that could occur by chance, P_o can overestimate reliability, especially when the categories are imbalanced. Despite this, it remains a useful baseline metric, particularly in exploratory or crowdsourced annotation studies. This measure provides an intuitive view of overall consistency and is often reported alongside chance-corrected statistics (Mohammad et al., 2018).

2.2. Bennett, Alpert, and Goldstein's S

Bennett, Alpert, and Goldstein's S coefficient provides a simple chance-corrected alternative to raw percentage agreement (BENNETT et al., 1954). It assumes that all categories are equally likely and is defined as

$$S = 1 - \frac{k(1 - P_o)}{k - 1},$$

where k is the number of possible categories. The S statistic adjusts for the probability of random agreement under the assumption of uniform category prevalence. However, it does not account for differences in annotator bias or imbalanced label distributions, which limits its suitability. Although rarely used in research, it is historically significant as one of the earliest attempts to correct for chance agreement metrics.

2.3. Cohen's Kappa

Cohen's κ (Cohen, 1960) is one of the earliest and most commonly used IAA measures for categorisation tasks. It applies when two annotators label the same set of items. κ compares the observed agreement P_o between the two annotators to the agreement expected by chance P_e, and is defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the proportion of items the annotators agreed on, and P_e is the proportion expected to agree by random chance (based on each annotator's labelling distribution). By subtracting P_e, κ corrects for random guessing. For example, if both annotators heavily favour one category, they might agree often simply by coincidence; κ will reduce the score to account for this. κ ranges from 1 (perfect agreement) down to 0 (no better than chance) or even negative (systematic disagreement). Cohen's paper introduced κ precisely to handle such situations where raw agreement is misleading.

2.4. Fleiss' Kappa

Fleiss' κ (Fleiss, 1971) generalises Cohen's κ to multiple annotators. It handles any number of annotators assigning nominal categories to a common set of items. Conceptually, Fleiss' κ also compares observed agreement to chance agreement, but it aggregates agreement across all annotators rather than comparing them pairwise. For each item i, the observed agreement among N annotators is computed as

$$P_i = \frac{1}{N(N-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1),$$

where n_{ij} is the number of annotators who assigned item i to category j. The mean observed agreement across all N items is then

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i.$$

The expected agreement P_e is derived from the overall category proportions, and the final κ is computed in the same way as Cohen's equation. Fleiss' κ assumes that each item receives the same number of annotations and can be sensitive to class imbalance or uneven marginal distributions. It is equivalent to Scott's π (Scott, 1955) under the condition that each item has the same number of annotations.

2.5. Weighted Kappa

Weighted kappa (Cohen, 1968) extends Cohen's κ to ordinal scales by assigning partial credit to near agreements. Instead of treating all disagreements equally, it applies a weighting function that penalises larger differences more heavily. The two most common weighting schemes are *linear*, where weights decrease proportionally with distance, and *quadratic*, where distant disagreements are penalised more strongly. Weighted κ is particularly useful for rating tasks on Likert or ordinal scales, as it captures both the direction and magnitude of disagreement between annotators.

2.6. Krippendorff's Alpha

Krippendorff's α (Krippendorff, 2013) is a versatile agreement metric that works for various data types (nominal, ordinal, interval, etc.), any number of annotators, and can handle missing data. α is popular in content analysis and increasingly in NLP because of this flexibility. It calculates agreement by focusing on disagreements:

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is the observed disagreement and D_e is the expected disagreement under chance. For nominal data, disagreement is usually defined as 0 if two labels are the same and 1 if different; for ordinal or interval data, a different distance function (e.g., squared difference) can be used. A major advantage of α is handling missing data. Notably, for nominal two-annotator cases with no missing values, α is mathematically equivalent to Cohen's κ; for multiple annotators it relates closely to Fleiss' κ. Researchers have used Krippendorff's α in NLP annotation of discourse or subjective content, where not all annotators label everything or where an ordinal scale needs to be handled appropriately.

2.7. Gwet's AC1/AC2

Gwet's AC1/AC2 (Gwet, 2001) are more recent chance-corrected agreement measures proposed to address some limitations of κ in situations of high agreement or highly skewed class distributions (often called the "κ paradox" where κ can be low despite high observed agreement, or vice versa). Gwet's AC1 (for nominal) and AC2 (for ordinal, with weights) use an alternative formula for expected agreement that tends to be more stable when classes are imbalanced or when annotators have bias. Recent NLP studies with extreme class imbalance have reported AC1 alongside κ to provide a more nuanced view of agreement (Chhun et al., 2024).

3. Structured annotations

Not all annotation tasks involve assigning a single label to an item. In many NLP tasks, annotators mark segments of text, which introduces alignment issues. Examples include named entity recognition (NER), where annotators highlight spans corresponding to entities, or text segmentation, where annotators divide a document into segments (e.g., topic boundaries or dialogue turns).

3.1. Span-Based Annotations

For span labelling tasks, a common approach is to treat one annotator's annotations as a *predicted* set of spans and another's as the *gold* set, then compute precision, recall, and F_1 score between them. This pairwise F_1 (or the Dice coefficient) measures overlap agreement, it rewards the annotators for each entity span they both identified exactly and penalises cases where one annotator found an entity that the other missed or where their span boundaries differ. In practice, F_1 is calculated on a per-span basis (exact matches).

3.2. Text Segmentation

When the task is to segment a text into contiguous units (e.g., dividing an article into sections), annotators are placing boundaries in text. Two popular metrics for segmentation agreement are:

- P_k (Beeferman et al., 1999): Slides a fixed-size window through the text and checks if the segmentations agree on whether there is a boundary between two points.
- WindowDiff (Pevzner and Hearst, 2002b): Similar to P_k, but it penalizes near-misses less harshly.

Producing a score between 0 and 1, where 0 indicates perfect agreement. If annotators' segment boundaries are slightly shifted, WindowDiff is more forgiving than P_k.

These metrics have been widely used in discourse and dialogue segmentation studies. For example, Scaiano and Inkpen (2012) applied them to evaluate segment boundaries in multi-party conversation, highlighting how boundary-based agreement can reveal both annotator consistency and the inherent ambiguity of conversational structure. Building on these approaches, Fournier and Inkpen (2012) introduced Segmentation Similarity, a generalised framework that unifies and extends P_k and WindowDiff by accounting for partial matches and variable boundary tolerance.

3.3. Unitising Tasks

In some tasks, annotators not only mark segments of text but also decide the number and positions of those segments, so-called unitising annotations. For example, in discourse analysis, annotators might break a conversation into discourse units and label each unit with a discourse function.

One metric is the holistic gamma (γ) measure (Mathet et al., 2015), which is designed for complex unitising tasks. Gamma treats the problem as a combination of segmentation and categorisation by finding an optimal alignment between units marked by different annotators and computing chance-corrected agreement. It accounts for both positional discrepancies (differences in unit boundaries) and categorical discrepancies (differences in labels).

3.4. Boundary Edit Distance

Boundary Edit Distance (Fournier, 2013) is a generalised measure of segmentation agreement that quantifies the minimal number of edits required to transform one annotator's segmentation into another's. It accounts for insertions, deletions, and near misses of boundaries, producing a flexible score that can handle varying degrees of segmentation granularity. Compared with WindowDiff or γ, Boundary Edit Distance is often considered a more robust alternative for tasks such as discourse or dialogue segmentation, where partial overlaps and fuzzy boundaries are common.

4. Continuous data

Certain tasks require continuous or interval-scale outputs rather than discrete categories. Examples include grading the fluency or coherence of text, or scoring emotional intensity on a numeric scale (Abercrombie et al., 2023; Wong and Paritosh, 2022). In such cases, the goal is to assess how consistently annotators assign scores along a continuous scale, rather than whether they select the same label. Measures of association or reliability for continuous data estimate how much of the variance in ratings reflects systematic differences between items rather than random noise or rater bias.

4.1. Intraclass Correlation Coefficient

The Intraclass Correlation Coefficient (ICC) is the most widely used reliability measure for continuous tasks. It quantifies the proportion of total variance in ratings that is attributable to true differences between items, relative to variance introduced by differences among raters or random error. Although well established in psychology and medicine, ICC is less frequently reported in NLP, where continuous annotation is often treated as ordinal or categorical.

Several ICC variants exist, distinguished by their model assumptions and by whether they estimate the consistency of individual ratings or the generalisability of aggregated ratings across multiple annotators. Following the framework introduced by Shrout and Fleiss (1979) and expanded by McGraw and Wong (1996a), the most common variants are:

- **ICC(1,1)**: One-way random effect, single measurement. Used when each item is rated by a different random set of raters.
- **ICC(1,k)**: One-way random effect, average measurement. It reflects the reliability of the mean rating when each item is rated by k raters.
- **ICC(2,1)**: Two-way random effect, single measurement (absolute agreement). Here, both items and raters are considered random samples from larger populations.
- **ICC(2,k)**: Two-way random effect, average measurement. It measures the reliability of the average rating from k raters.
- **ICC(3,1)**: Two-way mixed effect, single measurement (consistency). The raters are fixed (i.e., the particular raters are the only ones of interest) and the focus is on consistency rather than absolute agreement.
- **ICC(3,k)**: Two-way mixed effect, average measurement. It reflects the reliability of the average of k fixed raters.

In the two-way models, the difference between ICC(2,*) and ICC(3,*) lies in the treatment of the raters: in ICC(2,*) the raters are assumed to be randomly selected from a larger population, while in ICC(3,*) the raters are considered the only raters of interest (i.e., fixed). The choice between a single measurement (e.g., ICC(2,1)) and an average measurement (e.g., ICC(2,k)) depends on whether the study is concerned with the reliability of an individual rater's score or the aggregated score across raters. Selecting the appropriate ICC variant is essential for accurately assessing the reliability of annotations.

4.2. Cronbach's Alpha

Cronbach's α (Cronbach, 1951) is a measure of internal consistency that quantifies how closely related a set of ratings or items are as a group. It is mathematically equivalent to certain forms of the Intraclass Correlation Coefficient (ICC), specifically when the same raters assess each item under a one-way random-effects model. In annotation contexts, a high α suggests that annotators are applying the scale consistently across items. However, like ICC, its interpretation depends on model assumptions about rater and item effects, and it does not distinguish between consistency and absolute agreement.

4.3. Concordance Correlation Coefficient

The Concordance Correlation Coefficient (CCC) (Lawrence and Lin, 1989) assesses how well two sets of continuous ratings agree in both precision and accuracy. Unlike standard correlation coefficients, which measure only the strength of association, CCC also captures deviation from the identity line, that is, how close annotators' scores are to perfect concordance. It combines Pearson's correlation with a term penalising mean and scale differences between annotators. CCC has been widely used in affective computing and emotion intensity annotation, where it provides a more stringent criterion for agreement than correlation alone.

4.4. Correlation-Based Measures

Correlation coefficients assess the degree to which annotators produce similar rating patterns across items. They are useful for measuring consistency rather than absolute agreement.

For ordinal or ranked judgments, non-parametric correlation coefficients such as Spearman's ρ (Spearman, 1904) and Kendall's τ (Kendall, 1938) evaluate whether annotators preserve the same ordering of items, without assuming equal intervals between ranks. These measures are commonly used in human–model correlation studies, system ranking tasks, and as proxies for IAA when judgments are relative rather than absolute.

For continuous scores, Pearson's correlation coefficient (r) (Pearson, 1895) measures the strength and direction of the linear relationship between annotators' ratings. However, correlation alone does not indicate absolute agreement: two annotators may be perfectly correlated yet systematically offset in their scores. Therefore, correlation-based metrics should be interpreted as indicators of association rather than reliability in the strict sense.

5. Metric Selection and Interpretation

Selecting an appropriate IAA metric depends on the data type, number of annotators, and whether chance correction or missing data handling is essential. Table 1 summarises commonly used metrics, outlining their key properties and limitations to support informed metric selection and interpretation.

5.1. Interpretation of Agreement Scores

The qualitative interpretation of IAA scores varies across research domains and metric families. For κ-type coefficients such as Cohen's, Fleiss', and Krippendorff's α, conventions often follow the scale introduced by Landis and Koch (1977) and later refined by Viera et al. (2005) and McHugh (2012). These frameworks associate numerical ranges with descriptive categories such as "poor," "fair," "moderate," "substantial," and "almost perfect," though their thresholds are not universally accepted.

For reliability coefficients such as the ICC and Cronbach's α, psychometric and medical research typically applies more conservative standards, often considering values above 0.75 to indicate strong reliability (Koo and Li, 2016). Correlation-based measures, including Pearson's r, Spearman's ρ, and Kendall's τ, follow broadly similar conventions, distinguishing weak, moderate, and strong associations without fixed numeric boundaries.

Structured and segmentation-based metrics such as F_1, Boundary Similarity, Edit Distance, WindowDiff, P_k, and γ lack standard interpretive scales. Their scores depend heavily on task setup, annotation granularity, and tolerance for partial matches, making relative rather than absolute comparisons more informative.

Recent work has raised broader concerns about how IAA scores are interpreted and reported. Wong et al. (2021) argue that fixed interpretive thresholds are overly rigid for complex or subjective tasks and propose replication-based benchmarks in place of universal cutoffs. Klie et al. (2024) similarly show that many dataset papers report kappa-type coefficients without accounting for class imbalance, sample size, or annotator expertise, and call for greater transparency in reporting. Other studies highlight structural limitations: Stefanovitch and Piskorski (2023) find that traditional metrics often fail in multilingual or cross-domain settings, while Li et al. (2024a) demonstrate that categorical interpretive conventions do not directly transfer to structured or sequence-labelling tasks, where agreement is inherently lower due to hierarchical or contextual complexity.

Building on these critiques, Richie et al. (2022) emphasise that IAA should not be treated as a hard performance ceiling for machine learning systems, recommending instead the weighting or calibration of annotators rather than assuming uniform reliability. Wong and Paritosh (2022) introduce the k-rater reliability framework to quantify agreement across aggregated annotations, capturing stability that single-rater metrics often underestimate. Expanding this perspective, Klie et al. (2024) provide a meta-analysis of 591 NLP datasets, revealing inconsistent quality control and advocating for standardised, transparent reporting of annotation design and error estimation.

Each agreement metric is based on different underlying assumptions about how to measure consistency between annotators. Some account for chance agreement, while others measure overlap or correlation directly. Because of these differences, agreement scores are not always comparable across metrics and must be interpreted in relation to the task and data. Results can also be affected by factors such as class imbalance, the number of annotators, and annotation granularity. Agreement should therefore be understood as a context-dependent indicator of reliability rather than an absolute measure of annotation quality.

5.2. Reliability vs. Validity

Reporting confidence intervals alongside IAA metrics is essential, as they indicate the range within which the true value of a metric is likely to lie and quantify uncertainty in its estimation. Narrow intervals suggest high precision, while wider intervals reflect greater variability or instability (McHugh, 2012). Confidence intervals also enable more meaningful

Metric	Data Type	Missing Data?	Number of Annotators?	Chance-Corrected?	Sensitive to Imbalance?	Limitations
Percentage Agreement	Nominal	–	Pairwise	–	High	Overestimates reliability.
Bennett's S	Nominal	–	Two or more	✓	–	Assumes uniform label distribution.
Cohen's κ	Nominal	–	Two only	✓	High	Unstable when class frequencies or rater biases differ.
Fleiss' κ	Nominal	–	Three or more	✓	High	Requires equal numbers of ratings per item.
Krippendorff's α	Nominal / Ordinal / Interval / Ratio	✓	Two or more	✓	Moderate	Complex computation; interpretation can vary by distance metric.
Gwet's AC1/AC2	Nominal / Ordinal	✓	Two or more	✓	Low	Less common in literature; parameter selection affects results.
Weighted κ	Ordinal	–	Two only	✓	High	Requires careful choice of weighting scheme.
ICC	Continuous / Interval	✓	Two or more	✓	Moderate	Sensitive to model choice and assumes normality.
Cronbach's α	Continuous / Interval	✓	Two or more	✓	–	Assumes unidimensionality; may overestimate reliability.
CCC	Continuous	–	Two only	✓	Low	Sensitive to outliers; conflates accuracy and precision.
F_1 / Dice	Span-based / Structured	–	Pairwise	–	Task-dependent	Sensitive to boundary mismatches.
P_k / WindowDiff	Segmentation	–	Pairwise	–	–	Results dependent on window size.
Gamma (γ)	Unitising / Structured	✓	Two or more	✓	Moderate	Computationally intensive; conceptually complex.
Boundary Edit Distance	Segmentation	–	Pairwise	–	Task-dependent	Robust to near-misses; interpretation depends on tolerance.

Table 1: Overview of IAA metrics with key properties and limitations.

comparisons of agreement across tasks or groups by showing whether apparent differences are statistically significant or could arise from sampling variability (Reichenheim, 2004).

Although both p-values and confidence intervals support statistical inference, they serve different roles. A p-value assesses whether the observed agreement differs significantly from chance but provides no information about precision or effect magnitude. In contrast, a confidence interval conveys the range of plausible values for the agreement estimate, directly expressing its uncertainty. In IAA reporting, p-values can indicate significance, but confidence intervals offer a clearer sense of how stable and reliable the estimate is.

Inter-annotator agreement demonstrates reliability, the consistency with which annotators apply a scheme, but it does not establish validity. As Artstein and Poesio (2008) note, high IAA only confirms that annotators are consistent, not that they are measuring the intended construct correctly. High agreement can arise even from oversimplified or biased guidelines, while low agreement may reflect genuine ambiguity or interpretive diversity rather than poor data quality (Plank, 2022).

Researchers should therefore report reliability measures alongside evidence of validity, such as examples of ambiguous cases or comparisons to external references. Treating reliability and validity as complementary dimensions moves the field beyond numeric agreement scores and toward more interpretable and grounded evaluation practices (Howcroft et al., 2020). Achieving both requires clear task definitions, empirical grounding, and iterative refinement through pilot studies, adjudication, and expert feedback to distinguish methodological uncertainty from genuine subjectivity.

5.3. The Role of Disagreement

Disagreement among annotators is a common feature of both annotation and evaluation in NLP. Rather than treating all divergence as noise, it can reveal task ambiguity, underspecified guidelines, or variation in rater preferences. Prior work identifies multiple sources of disagreement, including genuine linguistic ambiguity, inconsistent criteria, and contextual or interface effects that influence decision-making (Reidsma and Carletta, 2008; Fleisig et al., 2023; Xu et al., 2024). These findings caution against interpreting low agreement as evidence of poor data quality.

Recent research increasingly views disagreement as informative rather than erroneous. Retaining label distributions or "soft labels" allows ambiguity to be represented explicitly, while measures of label dispersion indicate where annotators diverge (Rodríguez-Barroso et al., 2024; Fleisig et al., 2023). Rater-aware models extend this approach by learning annotator representations, disentangling systematic bias from item difficulty, and improving both aggregation and downstream robustness (Fleisig et al., 2023). Together, these methods treat disagreement as a meaningful signal about data, tasks, and annotator populations.

Practically, researchers are encouraged to report both agreement coefficients and measures of label dispersion, to analyse systematic disagreement across classes or subgroups, and to document annotation design factors such as training, and interface settings (Xu et al., 2024; Sandri et al., 2023). In subjective or interpretive tasks, preserving disagreement may be preferable to enforcing a single *ground truth*, as it better reflects population diversity and produces models that are more robust and representative (Rodríguez-Barroso et al., 2024; Sandri et al., 2023; Wan et al., 2023).

5.4. Effects of Pay and Time Pressure

Payment and time constraints influence both the quality of annotations and the behaviour of annotators. Early work showed that non-expert annotations can be valuable after aggregation, but quality varies with incentives and task design (Snow et al., 2008). Flat-rate payment schemes are common yet often inequitable, as completion times vary widely, leading to very low effective wages for slower or more careful annotators and incentivising speed over accuracy (Salminen et al., 2023). Performance-based or bonus systems can improve outcomes by promoting accuracy-oriented behaviour, though their effectiveness depends on task difficulty and the clarity of evaluation criteria (Rogstadius et al., 2011). Ethical analyses note that focusing solely on pay neglects broader issues of autonomy and bias, emphasising the need for transparency and worker agency (Shmueli et al., 2021). From a motivation perspective, excessive reliance on extrinsic rewards can displace intrinsic motives such as curiosity, though clear feedback and transparent incentive structures can help counteract this effect.

Time constraints have comparable consequences. Tight deadlines promote heuristic or superficial judgments and reduce exploration under uncertainty, potentially inflating agreement for the wrong reasons (Wu et al., 2022). Conversely, strict viewing-time limits reduce both performance and satisfaction, while extended sessions without breaks cause fatigue and increase variance in responses (Lim et al., 2024; Schlicher et al., 2021). Annotation difficulty also affects completion time, making uniform per-item pay problematic when item complexity varies (Wei et al., 2018). In practice, payment and timing protocols should prioritise accuracy over throughput: establish fair hourly rates based on observed completion times, define transparent evaluation criteria, avoid excessive time caps, and adjust compensation for more complex items. These design principles, reported alongside agreement statistics, support both reproducibility and ethical evaluation (Shmueli et al., 2021).

5.5. Human–Model Comparison

Large Language Models (LLMs) are now used not only as systems to be evaluated but also as evaluators, producing judgments of fluency, coherence, or factuality that were once the exclusive domain of human annotators. This shift challenges the long-held assumption that human agreement represents the upper bound of evaluation quality. Recent studies demonstrate that model-based evaluation can equal or even exceed human reliability on several text quality dimensions (Gilardi et al., 2023). As a result, human annotations can no longer serve as an unquestioned gold standard when models are expected to complement or surpass human performance.

Surveys have shown that the emergence of LLM-based evaluation has reshaped the methodological landscape of NLP (Li et al., 2024b; Gu et al., 2024; Chang et al., 2024; Laskar et al., 2024). These reviews highlight that while models often display higher internal consistency than human raters, they also risk reproducing systematic biases or calibration errors. Conversely, human disagreement can reflect genuine ambiguity or contextual sensitivity that models fail to capture. Rather than replacing human judgment, model-based evaluators should be benchmarked against diverse human perspectives and assessed for alignment with multiple criteria rather than correlation with a single reference set. Finally, Bojić et al. (2025) compare human annotators and several LLMs across sentiment, political leaning, emotional intensity, and sarcasm detection, finding that models often match human reliability on structured tasks but still underperform on nuanced or affective judgments, highlighting the continued value of human evaluation.

5.6. Annotator Expertise and Domain Knowledge

Annotator experience and subject-matter knowledge strongly affect both annotation quality and reliability. In knowledge-intensive domains such as legal text, domain experts are often essential, as they can apply specialised concepts consistently and resolve ambiguities that non-experts may misinterpret (Yang et al., 2019). Even among experts, however, variability persists, as shown in medical imaging and other high-stakes tasks, reflecting the inherent complexity of annotation in such settings (Yang et al., 2023). For general tasks such as sentiment analysis or entity tagging, aggregated non-expert labels can achieve high reliability (Zlabinger et al., 2020; Snow et al., 2008). Training and calibration also play a key role: novice annotators typically introduce higher variance, but consistency improves through structured feedback, curated examples, and repeated exposure (Huang et al., 2020; Zlabinger et al., 2020).

Expertise, however, extends beyond factual knowledge to include interpretive perspective. Homogeneous expert groups may achieve high agreement but risk amplifying shared biases and overlooking alternative viewpoints. Mixed-expertise or crowdsourced settings tend to be noisier but are valuable for subjective or contested phenomena such as offensiveness, subjectivity, or inference, where disagreement may reflect genuine diversity rather than error (Mehta and Srikumar, 2023; Snow et al., 2008). Accordingly, documenting annotator backgrounds, training procedures, and recruitment strategies is critical for contextualising agreement scores and assessing generalisability. Transparent reporting of such information supports both reproducibility and ethical accountability in annotation research (Wang et al., 2013).

Cultural and linguistic background also shapes annotation behaviour. Hershcovich et al. (2022) emphasise that NLP must account for cultural as well as linguistic variation, since interpretations of meaning and social norms differ across communities. Lee et al. (2024) report large cross-country differences in English hate-speech labels, and Pang et al. (2023) find similar variation across cultural groups in human-labelling tasks. Together, these studies show that annotation reflects cultural perspective as much as linguistic competence and highlight the importance of documenting annotator backgrounds and considering cross-cultural diversity in dataset design.

6. Conclusion

IAA is fundamental to ensuring the reliability and reproducibility of annotated data in NLP, but its value depends on more than just reporting a number. Reliable evaluation requires aligning the agreement measure with the task and rater design, clearly communicating underlying assumptions, and reporting uncertainty to reflect the limits of precision. Equally important is examining patterns of disagreement to understand whether they reflect ambiguity, bias, or inconsistency. Treating IAA as an integral part of the methodological process can lead to more transparent and interpretable annotation and evaluation practices.

7. Limitations

This paper aims to provide an accessible overview of IAA measures used across different types of NLP annotation and evaluation tasks. As such, it focuses on breadth rather than depth, outlining commonly applied metrics and their practical implications. Readers seeking detailed mathematical descriptions or domain-specific adaptations are encouraged to explore the primary literature cited throughout this paper. Practitioners are additionally advised to carefully plan their annotation task before committing to a specific metric.

While we have aimed to cover a wide range of IAA metrics, it is possible that certain tasks or niche metrics have been omitted. Areas such as multimodal annotation, interactive evaluation, and subjective phenomena like humour or creativity may require approaches beyond those discussed in this paper.

Acknowledgements

Joseph James is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

8. Bibliographical References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*.

Douglas G Altman. 1990. *Practical statistics for medical research*. Chapman and Hall/CRC.

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

Simone Balloccu, Anya Belz, Rudali Huidrom, Ehud Reiter, Joao Sedoc, and Craig Thomson, editors. 2024. *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

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

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

E. M. BENNETT, R. ALPERT, and A. C. GOLDSTEIN. 1954. Communications through limited-response questioning*. *Public Opinion Quarterly*, 18(3):303–308.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, 15(1):11477.

Alexander Braylan, Omar Alonso, and Matthew Lease. 2022. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *Proceedings of the ACM Web Conference 2022*, pages 1720–1730.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142.

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

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

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

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

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North Amer-*

ican Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161, Montréal, Canada. Association for Computational Linguistics.

Sayan Ghosh, Tejas Srinivasan, and Swabha Swayamdipta. 2024. Compare without despair: Reliable preference evaluation with generation separability. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12787–12805, Miami, Florida, USA. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Kilem Gwet. 2001. Handbook of inter-rater reliability. *Gaithersburg, MD: STATAXIS Publishing Company*, pages 223–246.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

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

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Ting-Hao'Kenneth' Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C Lee Giles. 2020. Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. *arXiv preprint arXiv:2005.02367*.

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

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Jacob Kohen. 1960. A coefficient of agreement for nominal scale. *Educ Psychol Meas*, 20:37–46.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*, 3rd edition. SAGE Publications.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Diya Li, Carolyn Rose, Ao Yuan, and Chunxiao Zhou. 2024a. Estimating agreement by chance for sequence annotation. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5085–5097, Bangkok, Thailand. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Gordon Lim, Stefan Larson, Yu Huang, and Kevin Leach. 2024. Towards fair pay and equal work: Imposing view time limits in crowdsourced image classification. *arXiv preprint arXiv:2412.00260*.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Kenneth O McGraw and Seok P Wong. 1996a. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30.

Kimberly O. McGraw and Sean P. Wong. 1996b. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Maitrey Mehta and Vivek Srikumar. 2023. Verifying annotation agreement without multiple experts: A case study with Gujarati SNACS. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10941–10958, Toronto, Canada. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Rock Yuren Pang, Jack Cenatempo, Franklyn Graham, Bridgette Kuehn, Maddy Whisenant, Portia Botchway, Katie Stone Perez, and Allison Koenecke. 2023. Auditing cross-cultural consistency of human-annotated labels for recommendation systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1531–1552.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.

Leonid Pevzner and Marti Hearst. 2002a. A critique and improvement of the p_k metric for text segmentation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 27–36. Association for Computational Linguistics.

Lev Pevzner and Marti A Hearst. 2002b. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

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

Michael E Reichenheim. 2004. Confidence intervals for the kappa statistic. *The Stata Journal*, 4(4):421–428.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Russell Richie, Sachin Grover, and Fuchiang (Rich) Tsui. 2022. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 275–284, Dublin, Ireland. Association for Computational Linguistics.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M. Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the international*

AAAI conference on web and social media, volume 5, pages 321–328.

Joni Salminen, Ahmed Mohamed Sayed Kamel, Soon-Gyo Jung, Mekhail Mustak, and Bernard J Jansen. 2023. Fair compensation of crowdsourcing work: the problem of flat rates. *Behaviour & Information Technology*, 42(16):2871–2892.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Martin Scaiano and Diana Inkpen. 2012. Getting more from segmentation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 362–366, Montréal, Canada. Association for Computational Linguistics.

Katharina D Schlicher, Julian Schulte, Mareike Reimann, and Günter W Maier. 2021. Flexible, self-determined... and unhealthy? an empirical study on somatic health among crowdworkers. *Frontiers in Psychology*, 12:724966.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Nicolas Stefanovitch and Jakub Piskorski. 2023. Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 71–86, Singapore. Association for Computational Linguistics.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, 7(1):258.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.

Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2018. Clinical text annotation–what factors are associated with the cost of time? In *AMIA Annual Symposium Proceedings*, volume 2018, page 1552.

Ka Wong and Praveen Paritosh. 2022. k-Rater Reliability: The correct unit of reliability for aggregated human annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–384, Dublin, Ireland. Association for Computational Linguistics.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.

Charley M Wu, Eric Schulz, Timothy J Pleskac, and Maarten Speekenbrink. 2022. Time pressure changes how people explore and respond to uncertainty. *Scientific reports*, 12(1):4122.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

Feng Yang, Ghada Zamzmi, Sandeep Angara, Sivaramakrishnan Rajaraman, André Aquilina, Zhiyun Xue, Stefan Jaeger, Emmanouil Papagiannakis, and Sameer K Antani. 2023. Assessing inter-annotator agreement for medical image segmentation. *IEEE Access*, 11:21300–21312.

Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C Wallace, and Ani Nenkova. 2019. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. *arXiv preprint arXiv:1905.07791*.

Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, Mete Sertkan, and Allan Hanbury. 2020. Dexa: supporting non-expert annotators with dynamic examples from experts. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2109–2112.