Can We Hide Machines in the Crowd? Quantifying Equivalence in LLM-in-the-loop Annotation Tasks

Jiaman He
RMIT University
Naarm/Melbourne, Australia
jiaman.he@student.rmit.edu.au

Zikang Leng
Georgia Institute of Technology
Atlanta, USA
zleng7@gatech.edu

Dana McKay
RMIT University
Naarm/Melbourne, Australia
dana.mckay@rmit.edu.au

Damiano Spina
RMIT University
Naarm/Melbourne, Australia
damiano.spina@rmit.edu.au

Johanne R. Trippas
RMIT University
Naarm/Melbourne, Australia
j.trippas@rmit.edu.au

Abstract

Many evaluations of large language models (LLMs) in text annotation focus primarily on the correctness of the output, typically comparing model-generated labels to human-annotated "ground truth" using standard performance metrics. In contrast, our study moves beyond effectiveness alone. We aim to explore how labeling decisions–by both humans and LLMs–can be statistically evaluated across individuals. Rather than treating LLMs purely as annotation systems, we approach LLMs as an alternative annotation mechanism that may be capable of mimicking the subjective judgments made by humans. To assess this, we develop a statistical evaluation method based on Krippendorff's α, paired bootstrapping, and the Two One-Sided t-Tests (TOST) equivalence test procedure. This evaluation method tests whether an LLM can blend into a group of human annotators without being distinguishable. We apply this approach to two datasets, MovieLens 100K and PolitiFact, and find that the LLM is statistically indistinguishable from a human annotator in MovieLens 100K ($p = 0.004$), but not in PolitiFact ($p = 0.155$), highlighting task-dependent differences.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

Keywords

Text Annotation, LLM Evaluation, Validity of Experimentation

ACM Reference Format:
Jiaman He, Zikang Leng, Dana McKay, Damiano Spina, and Johanne R. Trippas. 2025. Can We Hide Machines in the Crowd? Quantifying Equivalence in LLM-in-the-loop Annotation Tasks. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2025), December 7–10, 2025, Xi'an, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3767695.3769508

SIGIR-AP 2025, Xi'an, China

ACM ISBN 979-8-4007-2218-9/2025/12
https://doi.org/10.1145/3767695.3769508

1 Introduction

In Alan Turing's 1950 landmark 1950 paper, he proposed a criterion for machine intelligence: if a machine can engage in conversation such that a human evaluator cannot reliably distinguish it from another human. The machine can be said to exhibit intelligent behaviour [53]. This formulation, now known as the Turing Test, shifts the focus from a machine's internal workings to its external behaviour. In this work, we build on the spirit of the Turing Test. However, we focus on a more domain-specific setting: text annotation tasks. Instead of testing a machine through open-ended conversations, we ask a simpler question. Can an LLM act like a human annotator? Specifically, can its output be statistically indistinguishable from that of people in a multi-person annotation task?

While LLMs have shown strong performance on many general-purpose classification tasks [22, 38, 56], their role in subjective or domain-specific annotation settings remains unclear. In information retrieval contexts—such as document relevance assessment, intent classification, or stance detection—annotation often involves subtle, context-dependent judgments [17, 29, 43, 48]. These applications continue to rely on traditional classifiers trained on human-labeled data, particularly where interpretability, auditability, or fairness are required [31, 32]. If LLMs can produce labels that are indistinguishable from those generated by humans in such scenarios, they offer a path to reducing annotation costs while preserving human-like interpretive behavior. This makes it crucial to understand whether LLMs can accurately and effectively participate in the human annotation process used to train these models.

Text classification is widely used for natural language processing (NLP) tasks, such as news categorization, sentiment analysis, or subject labeling [15, 49]. It involves assigning labels to textual elements such as sentences, questions, paragraphs, or entire documents. Some of the classifications require human labeling to train and test the machine learning models. Annotations act as the ground truth against which models are tested, refined, and advanced.

Annotation tasks often depend on subjective human judgment rather than a single, objective truth [33]. For example, in relevance assessments, people may interpret the same content differently, and their judgments can change over time [17]. So, the aim is not always to identify the "correct" label, but to understand how labeling decisions emerge across different individuals. LLMs are increasingly good at producing convincing outputs. However, it is still unclear whether their responses reflect human judgment,

especially in the absence of evidence that their outputs are rooted in actual human experience [17].

Traditional human annotation can be constrained by cost and consistency [12]. Recent research has compared the quality of annotations by LLM and humans for NLP applications [38]. Also, a study investigated using LLMs to help or even replace human annotators on some tasks [1], usually by comparing their agreement with human results using measures like Krippendorff's α or Cohen's kappa. Most of these studies treat the LLM as a single, standalone system and check how well it matches a human-created "ground truth" [1, 6, 14, 32, 55].

We introduce an evaluation method for LLMs based on group dynamics. Rather than evaluating a model in isolation, we assess whether it can substitute a human within an annotation group without significantly altering the group's behavior. An LLM judgment is deemed successful if the LLM's presence is statistically indistinguishable from a human's presence.

This work treats LLMs not just as tools for text classification, but as participants that can imitate the subjective and sometimes inconsistent judgments made by humans. We introduce a practical method based Krippendorff's α, bootstrapping and TOST to test whether an LLM can blend into a group of human annotators without being identified. This approach requires only a small number of annotation items and functions as a domain-specific adaptation of the Turing Test. It supports early-stage evaluation on a small sample to determine the suitability of LLMs for large-scale annotation. We apply it to a real-world classification task and examine the results. Our key contributions are:

- We propose an evaluation methodology that statistically tests whether an LLM can substitute for a human annotator in multi-annotator text classification tasks.
- We demonstrate the application of our methodology on two datasets—MovieLens 100K and PolitiFact—showing that the LLM is statistically indistinguishable from human annotators in MovieLens 100K ($p = 0.004$) but not in PolitiFact ($p = 0.155$), revealing important task-dependent differences.
- We release a dataset containing LLM annotations alongside human annotations for a multi-annotator task. The dataset is publicly available at: https://github.com/peanutH/LLM-evaluation.

2 Related Work

2.1 Text Classification

Text classification is a core task in NLP. It helps organize unstructured text from sources like messages, documents, and websites. To make sense of all this text, researchers and developers use models that automatically classify text into categories such as topic, sentiment, or author identity [49]. In addition, labeled data is essential for building and testing models, with human annotations serving as the "ground truth" [38, 49]. Manual labeling is often costly, slow, and inconsistent, making automation preferable. Models are typically trained on a subset of labeled data and evaluated by comparing their predictions to human annotations.

Recently, LLMs have shown impressive performance across many NLP tasks [8], raising the possibility of replacing traditional supervised models or even human annotators. However, this shift is far from complete [37]. Many domain-specific real-world applications, such as medical coding, legal triage, or sentiment analysis, continue to rely on traditional classifiers trained on large datasets labeled by experts. This is because LLMs alone often lack the nuanced understanding required for these tasks. As a result, substantial human annotation remains essential for developing and validating reliable models. In areas where complex human judgment is critical, direct human involvement remains indispensable [9].

2.2 Human Annotation and Generative AI

Human annotation comes with challenges, most notably cost and consistency [12]. Text classification models rely on large amounts of labeled data, and hiring workers to do all the labeling is not always feasible. That is why crowdsourcing has become a solution. Platforms that connect requesters with crowd workers make it possible to outsource labeling tasks, such as relevance judgments, sentiment tagging, and topic categorization, to non-experts [2, 29]. This approach has been crucial in building datasets needed to train machine learning models.

Crowdsourcing was once viewed as a flexible and empowering option for workers. However, it is often criticized as invisible, low-paid labor that supports modern AI behind the scenes. Given these concerns, researchers have started asking: Can generative AI (GenAI), especially LLMs, step in and take over some of these annotation tasks? Some early findings suggest that LLMs tend to perform well on straightforward tasks, such as summarization or basic sentiment analysis [9]. But when the task requires more nuanced judgment—like interpreting sarcasm, ambiguity, or subtle context—human annotators still outperform the machines [38].

So, the real question is not just whether GenAI can get the "right" answer. It is whether its decisions reflect the kinds of judgments humans would make, especially in cases where there is no single correct label. One study [17] examined the ability of LLMs to handle relevance judgments, a task where subjectivity plays a role. While LLMs showed some ability to mimic human responses, they were not consistent or nuanced enough to fully replace human workers.

That is why we are developing a new evaluation framework to better understand whether LLM-generated annotations can be mistaken for human ones, not just in terms of correctness but in how closely they match human reasoning and subjectivity.

2.3 Existing Evaluation Method

Traditional evaluation methods for generative AI in annotation tasks typically benchmark AI-generated outputs against human annotations using metrics such as accuracy, precision, Kendall's τ, or inter-annotator agreement scores like Cohen's kappa and Krippendorff's α [1, 17, 55]. The approaches used in existing studies treat generative AI as a system and evaluate their output against the overall consensus of a crowd. In doing so, they prioritize alignment with collective human judgments, rather than examining how closely AI aligns with the characteristics of individual annotators.

Although these evaluations show whether generative models can produce generally accurate labels, they often miss how humans actually annotate. In real tasks—especially subjective ones

like relevance, sentiment, or moderation—judgments vary with expertise, interpretation, or background [35]. Treating this diversity as a single "gold standard" can overstate model capability [17].

Correlation metrics like Kendall's τ [26] are good for checking if LLM rankings match system-level outcomes. But they don't show how well LLMs fit into the social side of annotation, where disagreement and variation are normal. Measures like Cohen's κ [10] and Krippendorff's α [27] better capture consistency, but they still treat LLMs as outsiders compared to human annotators, rather than as active collaborators in the process.

This framing can lead to an inflated sense of how interchangeable LLMs are with humans, particularly in complex or cognitively demanding annotation settings [39]. As prior studies have noted, even as LLMs improve at mimicking human language and surface-level judgment, it remains a significant leap to assume their outputs are equivalent to human reasoning without verification. At present, there is no definitive evidence that LLM-generated judgments are grounded in human experience, intuition, or context.

This raises an important question: if an LLM's annotation looks like a human's, does that mean it is truly the same, or are we missing differences in how judgments are made? In many tasks, there is no single "correct" answer; human judgments are often subjective, context-dependent, and inconsistent over time [5, 17]. To better reflect this, we propose a new approach: instead of checking if an LLM agrees with the crowd, we ask whether it can blend into the crowd—becoming statistically indistinguishable from human annotators, and that we call it a "Statistical Turing Test".

2.4 Human Judgment

Before comparing LLMs to humans, it is important to first understand the nature of human judgment. Human judgment is commonly modeled as a cognitive process that aligns well with linear models of cue integration [19, 23]. In such models, people make decisions based on a set of cues, each weighted differently depending on its perceived importance.

Brehmer et al. [7] noted that linear models tend to fit human judgments quite well. Even when nonlinear or configural components are present, they typically account for only a small portion of the variance, and their generalizability across tasks is uncertain. Additionally, human judgments are often inconsistent, with the level of consistency varying according to the predictability of the task. There are also substantial inter-individual differences in how people weigh signals, even among individuals with considerable experience on the task.

This inconsistency can be attributed to the variability in cue weights applied across different tasks. Prior research has shown that judgment consistency tends to decrease as the number of cues increases [16]. In contrast, LLMs often display greater consistency in annotation tasks [17], likely due to more stable internal representations of cues and weights. To evaluate whether the LLM's cue integration falls within an acceptable threshold of variability, we use inter-annotator agreement (IAA) as a proxy for measuring consistency in annotation judgments.

2.5 Inter-Annotator Agreement (IAA)

Researchers who rely on hand-labeled data, where items are manually labeled with categories for empirical analysis or model development, must demonstrate that the labeling process is reliable [4]. A fundamental assumption in annotation methodology is that the data are considered reliable when multiple annotators agree on the labels assigned, to a degree appropriate for the objectives of the study [11, 28].

Consistent agreement among annotators suggests that they share a common understanding of the annotation guidelines, and thus can be expected to apply those guidelines consistently. IAA is a metric used to quantify this consistency. In multi-people annotation settings, annotators may have varied backgrounds and limited domain expertise. IAA helps determine whether labels are trustworthy and whether a task is clearly defined or inherently subjective. High agreement indicates clear instructions and straightforward data, while low agreement may reveal task ambiguity, multiple valid interpretations, or inconsistent annotator behavior.

Beyond assessing label quality, IAA also serves as a diagnostic tool for identifying issues in the annotation process. By examining patterns of agreement and disagreement, researchers can uncover sources of ambiguity, identify annotator bias, and refine the guidelines. One of the most used IAA measure is Krippendorff's α.

2.6 Krippendorff's α

Krippendorff's α is a robust and widely-used reliability coefficient for measuring inter-annotator agreement, particularly when annotations are incomplete, involve more than two coders, or span different levels of measurement (nominal, ordinal, interval, etc.) [27]. Unlike simpler metrics such as Cohen's kappa, which assume fixed pairwise comparisons, Krippendorff's α can accommodate complex and realistic annotation setups—including crowdsourced data with missing entries or unequal contributions from annotators.

Krippendorff's α quantifies the extent to which observed disagreement differs from what would be expected by chance, with values ranging from 1 (perfect agreement) to 0 (chance-level agreement) and negative values indicating systematic disagreement. Importantly, Krippendorff's α is sensitive not only to consistency but also to the nature of the scale being used, making it well-suited for subjective or ambiguous tasks where subtle distinctions matter.

In our evaluation framework, we employ Krippendorff's α to assess whether annotations produced by an LLM achieve a comparable level of agreement with human annotators as humans achieve with one another. Rather than simply comparing the LLM to a gold standard, we integrate it into the annotator pool and compute Krippendorff's α across the mixed group.

3 Methodology

Building on the idea discussed in Section 2.5, consistent labeling by multiple human annotators suggests that they share a common understanding of the annotation guidelines and apply them reliably [4]. Inspired by the logic of the Turing Test [53], we propose a methodology that uses inter-annotator agreement to evaluate whether an LLM can function as an individual annotator. That is, if it can serve as a substitute for a human in the annotation process.

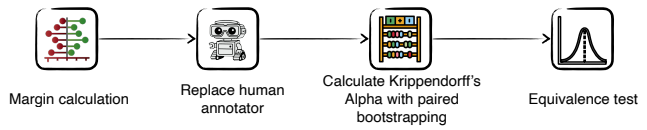

Figure 1: Our evaluation methodology workflow

In this section, we outline the methodology used to evaluate whether an LLM can effectively substitute for a human annotator by examining changes in inter-annotator agreement, the methodology workflow is shown in Figure 1. We first introduce the rationale behind this evaluation (Section 3.1.1). We then detail the protocol for substituting a human annotator with an LLM (Section 3.1.2), followed by the approach for measuring how agreement levels vary due to these substitutions (Section 3.1.3). Next, we describe our statistical procedures for estimating variability in agreement scores using a paired bootstrap method (Section 3.1.4) and establishing equivalence through Two One-Sided Tests (TOST) (Section 3.2). Finally, we discuss practical considerations for determining appropriate sample sizes and annotator group sizes to ensure robust and reliable results (Section 3.3).

3.1 LLM Substitution Protocol

3.1.1 Motivation: Substituting Human Annotators. Krippendorff's α measures the extent of agreement among annotators using the formula:

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

where D_o is the observed disagreement and D_e is the expected disagreement by chance.

To understand how α behaves under substitution, consider a group of three annotators-A, B, and C—whose annotations yield an agreement score α_1. Now imagine replacing annotator A with a new annotator E and computing a new α value, α_2. If $\alpha_1 \approx \alpha_2$, this suggests that annotator E exhibits a similar consistency pattern to annotator A for the same task. Repeating this comparison with different annotators and observing small differences (i.e., $|\alpha_1 - \alpha_2|$ within a tolerable range) may indicate that the consistency patterns among the annotators are comparable.

This idea motivates our approach: if an LLM can replace a human annotator without significantly altering the inter-annotator agreement, it may be acting as a reasonable substitute.

3.1.2 Protocol: Replacing Annotators with an LLM. To formalize this idea, we consider a group of i human annotators who have independently labeled a shared set of n items. We then simulate the substitution process by replacing one human annotator at a time with the LLM. This results in i modified annotation groups—each with $i - 1$ humans and one LLM.

In each iteration, we remove the annotations from one human (e.g., annotator A) and replace them with labels generated by the LLM for the same items. The LLM effectively stands in for the removed annotator. This process is repeated for all i human annotators. The substitution progression is illustrated in Figure 2.

In cases where the original annotator did not label all items, we only substitute the entries that exist—i.e., the LLM only replaces

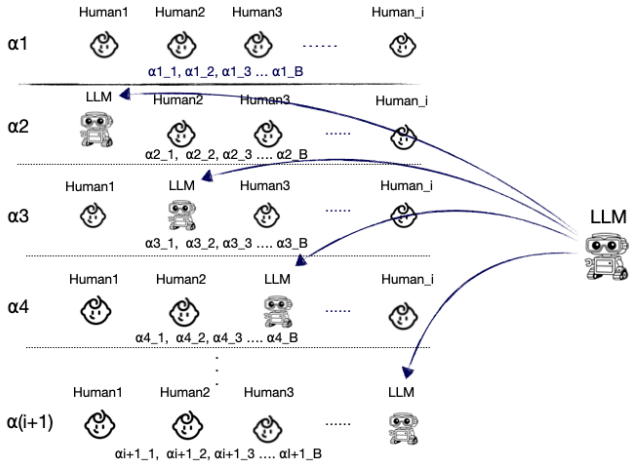

Figure 2: LLM substitution

the ratings for items the original annotator labeled. Items left blank by the original annotator remain blank.

3.1.3 Measuring Agreement Across Substitutions. In complex annotation tasks involving comprehension, reasoning, or subtle interpretation, perfect agreement is unlikely—even among humans. We therefore don't expect the LLM to match human annotations exactly. Instead, we evaluate whether the LLM can match the overall consistency of human annotators.

We begin by computing Krippendorff's α for the original group of i human annotators. We denote this baseline agreement as:

$$\alpha_1 = \text{Krippendorff's } \alpha \text{ for human group}$$

Next, we compute α for each of the i modified groups where one human annotator is replaced by the LLM. These are denoted as:

$$\alpha_j = \text{Krippendorff's } \alpha \text{ for substitution group } j = 2, \ldots, i+1$$

By comparing the LLM-substituted scores α_j with the original human score α_1, we can observe how agreement changes when the LLM replaces a human annotator in the annotation group.

3.1.4 Paired Bootstrap for Variability Estimation. To assess the variability of agreement scores and ensure that observed differences are statistically meaningful, we apply the paired bootstrap method inspired by Krippendorff [27]

Prior work has shown that analyzing a subset comprising just 40% of the full dataset can provide a reliable estimates of inter-rater agreement [3]. Following this insight, we apply bootstrap sampling [52] to resample the annotation data.

Specifically, we perfrom B bootstrap iterations (eg., $B = 300$), where in each iteration we sample N items with repalcement from the full set of n annotated items. For each sample, we compute the Krippendorff's α.

Let $\alpha_1^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_1^{(B)}$ represent the α values computed for the original human group across the B bootstrap samples.

$$\alpha_1 = \left\{\alpha_1^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_1^{(B)}\right\} \quad (1)$$

For each LLM-substituted group $j = 2, \ldots, i+1$, we similarly compute a distribution of alpha values:

$$\alpha_j = \left\{ \alpha_j^{(1)}, \alpha_j^{(2)}, \ldots, \alpha_j^{(B)} \right\} \tag{2}$$

To ensure fair comparison, we use the same bootstrap samples (i.e., same sampled items) across all groups in each iteration. This paired bootstrap procedure allows us to compare the variability in agreement across human and LLM-substituted groups under consistent sampling conditions.

3.2 Equivalence Testing with TOST

We then summarize the results by computing the mean agreement score for the original human–human annotations:

$$x_2 = \frac{1}{B} \sum_{f=1}^{B} \alpha_1^{(f)}$$

and the average agreement score across all LLM–human replacement cases:

$$x_1 = \frac{1}{i \cdot B} \sum_{j=2}^{i+1} \sum_{f=1}^{B} \alpha_j^{(f)}$$

Here, x_1 represents the overall mean of $\alpha_2, \alpha_3, \ldots, \alpha_{i+1}$, capturing average agreement when one human is replaced by the LLM. x_2 represents the baseline agreement among all-human groups.

These means serve as the basis for an equivalence test using the Two One-Sided t-Tests (TOST) procedure [42, 46, 50]. The goal is to determine whether the difference $|x_1 - x_2|$ falls within a pre-defined equivalence margin Δ_{equiv}, which indicates a practically negligible difference in reliability.

We compute the two TOST statistics as follows:

$$t_1 = \frac{(x_1 - x_2 - \Delta_{\text{equiv}})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{and} \quad t_2 = \frac{(x_1 - x_2 + \Delta_{\text{equiv}})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s is the pooled standard deviation of the two sets ofαscores, and B_1, B_2 are the sample sizes of the two groups.

The null hypothesis is that the difference exceeds the equivalence margin:

$$H_0 : |x_1 - x_2| > \Delta_{\text{equiv}}$$

We reject H_0 if both t_1 and t_2 fall within the critical region for their respective one-sided tests, thereby concluding that the observed difference is within an acceptable range of equivalence.

3.2.1 Equivalence Margin Definition. The equivalence margin Δ_{equiv} is a threshold below which the difference is considered negligible. We estimate this margin empirically based on natural variability among human annotators:

$$\Delta_{\text{equiv}} = (\alpha_a - \alpha_b) \cdot \text{fraction}$$

where:

- α_a is Krippendorff's α calculated from a group of human annotators (e.g., annotators 1–3),
- α_b is Krippendorff's α from another independent group (e.g., annotators 4–6),
- Both groups annotate the same items using the same guidelines,
- $\alpha_a - \alpha_b$ reflects typical human-to-human variability,
- `fraction` is a scaling factor (e.g., 0.5 or 0.8) that controls how strict the equivalence test is — smaller values require the LLM to match humans more closely.

By grounding the margin in actual human variability, this approach makes the equivalence test realistic and interpretable. The margin reflects what is already tolerated in human-human comparisons, rather than relying on arbitrary cut-offs.

3.3 Sample Size and Annotator Group Size

3.3.1 Dataset Size. To determine an appropriate minimum bootstrap sample size for calculating Krippendorff's alpha, we follow Bloch and Kraemer's formula [28], which takes into account the desired minimum agreement level $\alpha_{\min}$, a confidence level z, and the probability of observing agreement by chance p_c:

$$N = z^2 \left(\frac{(1 + \alpha_{\min})(3 - \alpha_{\min})}{4(1 - \alpha_{\min})p_c(1 - p_c)} \right)$$

Following the paired bootstrap procedure outlined in Section 3.1.4, we randomly sample N items from the dataset in each bootstrap iteration. Using $z = 0.95$, $\alpha_{\min} = 0.8$, and $p_c = 0.17$, we calculate the minimum required size of each bootstrap sample to be $N = 32$. As noted in Section 3.1.4, a bootstrap sample comprising 40% of the full dataset is sufficient to yield a reliable estimate of interrater agreement. This implies that the full dataset should contain at least $n = 2.5N = 80$ items.

3.3.2 Annotator Group Size. We analyze how Krippendorff's α changes when one human annotator is replaced by an LLM. Let i be the number of annotators, n the number of items, and ℓ_{ak} the label from annotator a on item k. As mentioned in Section 2.6, Krippendorff's α is defined as:

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

where D_o is the average observed pairwise disagreement and D_e is the expected disagreement under random labeling based on marginal label distributions.

Change in observed disagreement. When coder r is replaced by an LLM that assigns labels L_k, the change in D_o becomes:

$$\Delta D_o = \frac{2}{n\,i} \sum_{k=1}^{n} \left(\bar{d}_{\text{LLM},k} - \bar{d}_{r,k} \right),$$

where $\bar{d}_{\text{LLM},k}$ and $\bar{d}_{r,k}$ denote the average disagreement between the LLM (or original coder r) and all other annotators on item k.

Change in expected disagreement. Substituting in the LLM slightly alters the marginal distribution of labels, shifting $p_c \mapsto p_c' = p_c + \Delta p_c$. The first-order change in D_e is then:

$$\Delta D_e \approx -2 \sum_{c} p_c \Delta p_c.$$

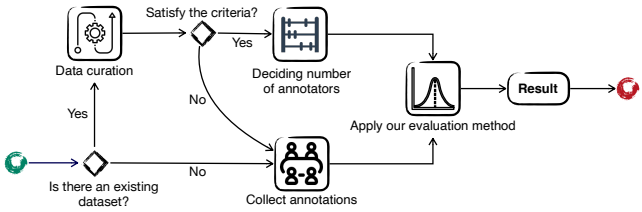

Figure 3: Guideline for applying our methodology

Table 1: Datasets considered for evaluation.

Dataset	Domain	# Items
MovieLens 100K [20]	Movie ratings	1,682
WebCrowd25k [30]	IR relevance	≈4,500
TREC-8 Re-assessments [44]	IR relevance	4,269
Familiarity–QuerySpec [21]	Query specificity classification	83
MBIC [51]	Media bias	≈2,600
CoQA [41]	Conversational QA	8,000
POPQUORN [40]	Offensive–QA	≈5,500
D3CODE [36]	Cross-cultural offensiveness detection	≈4,500
CrowdsourcingTruthfulness–PolitiFact [45]	Misinformation – Veracity classification	120
HateXplain [34]	Hate speech detection	≈20 000
CODA-19 [24]	COVID-19 abstract section labeling	10,966

Total change in α. Using a first-order Taylor expansion of α, we obtain:

$$\Delta\alpha \approx -\frac{1}{D_e}\Delta D_o + \frac{D_o}{D_e^2}\Delta D_e.$$

Substituting the expressions above yields:

$$\Delta\alpha \approx -\frac{2}{n\,i\,D_e}\sum_{k=1}^{n}\left(\bar{d}_{\mathrm{LLM},k} - \bar{d}_{r,k}\right) \;-\; \frac{2\,D_o}{D_e^2}\sum_{c} p_c\,\Delta p_c.$$

The derivation[1] shows that $\Delta\alpha$, resulting from substituting one human annotator with an LLM, is inversely proportional to the number of annotators i (i.e., $\Delta\alpha \propto \frac{1}{i}$). This implies that as the number of annotators increases, the impact of a single substitution on α becomes smaller. To select an appropriate group size, we identify the elbow point on the curve of $\Delta\alpha$ versus the number of annotators, where the rate of change drops sharply. We apply the L-method [47] to detect this point by fitting two lines to the curve—one before and one after each candidate split—and choosing the split that minimizes the total fitting error. This method captures the transition from rapid to gradual change.

4 Experimental Evaluation

To validate our evaluation method, we designed an experimental workflow, illustrated in Figure 3. Since our experiments are conducted on existing datasets, we follow the corresponding branch of the workflow.

4.1 Dataset

We compiled a list of publicly available datasets that provide multiple annotations per item along with identifiable annotator IDs, enabling us to trace which individual labeled each item and to perform annotator-level substitutions. These datasets are summarized in Table 1. To fit our evaluation experiments, we filtered the datasets using a set of selection criteria defined by our evaluation methodology. The code for this filtering process is available.[2] Specifically, we required that the dataset contain two disjoint groups of annotators, each of size i, such that there are at least 80 items, with each item annotated by at least two annotators from each group. After data filtering, two datasets satisfy our criteria: MovieLens 100K [20] and PolitiFact [45].

The MovieLens 100K dataset, collected by the GroupLens Research Project at the University of Minnesota [20], contains 100,000 movie ratings (on a 1–5 scale) from 943 users across 1,682 movies. Each user rated at least 20 movies. The dataset also includes basic demographic information about the users, such as age, gender, occupation, and ZIP code.

The original PolitiFact dataset collected by Wang [54] contains 12,000 statements produced by U.S. politicians, each statement is labeled by an expert judge on a six-level scale for the statement's truthfulness. Roitero et al. [45] selected 120 statements, 20 for each truth level, related to COVID-19 from the original PolitiFact dataset. Then, workers were recruited from Amazon Mechanical Turk to annotate each statement. Overall, each statement was annotated by 10 workers over three different scales: three-level, six-level, one-hundred level. In this work, we use the three-level dataset.

To obtain the LLM annotations, we used GPT-4o mini to annotate the same set of items selected for this experiment, as described in Section 4.2. We designed the prompt, see Figure 4, to follow the same annotation guidelines provided to human annotators.

> –SYSTEM–
> You are an average movie watcher. Rate each movie from 1 to 5 based on how much you liked it overall. Consider the story, acting, and overall enjoyment.
>
> –USER–
> Respond with a list of ratings, one for each movie, in the same order as presented. Only include the numeric ratings and nothing else.
>
> Here are the movies for you to rate:
> <LIST_OF_MOVIES>

Figure 4: LLM prompt for generating annotations for the MovieLens dataset. Further information is available on the paper's GitHub.

4.2 Data Selection

MovieLens 100k. For evaluation, we selected 38 coders split into two groups of 19 coders each. We used the elbow point finding method detailed in Section 3.3 to determine the optimal number of annotators in a group. As shown in Figure 5, this point was at 19 annotators per group. We divided the coders into two groups:

- **Group A:** Human_1, to Human_{19}
- **Group B:** Human_{20} to Human_{38}

[1] Full derivation can be found in https://github.com/peanutH/LLM-evaluation

[2] Our filtering code can be found in https://github.com/peanutH/LLM-evaluation

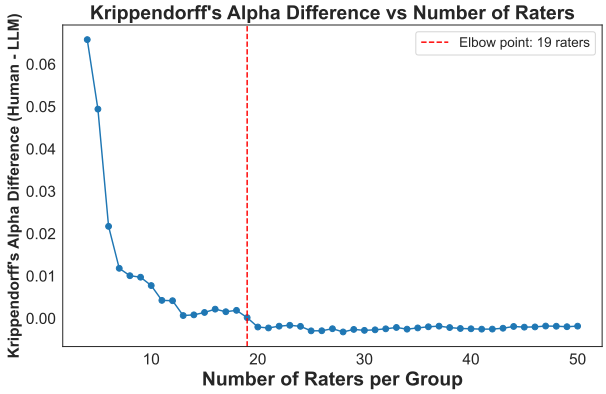

Figure 5: p-values obtained for the MovieLens 100K dataset for various value of i, the number of annotators in the group.

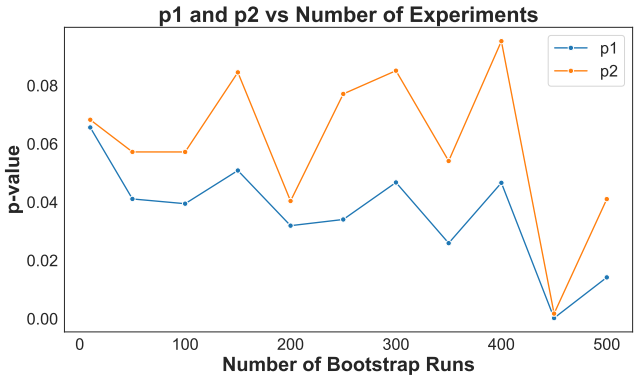

Figure 6: p-values obtained for the MovieLens 100K dataset for various value of B, the number of bootstrap iterations.

CrowdsourcingTruthfulness–PolitiFact. We selected 86 coders for the evaluation—the minimum number that could be filtered from the dataset while still satisfying our evaluation criteria. The coders were divided evenly into two groups:

- **Group A:** Human_1 to Human_{43}
- **Group B:** Human_{44} to Human_{86}

This grouping enables the calculation of the equivalence margin (see Section 3.2.1). To satisfy the requirements for applying Krippendorff's α that each item must be annotated by at least two coders within a group, and each coder must annotate at least one item [28], we filtered a subset of 100 items from the MovieLens 100k dataset and the PolitiFact dataset, each annotated by the selected coders. This meets the minimal number of items discussed in Section 3.3.

4.3 LLM Substitution Evaluation

We evaluated whether LLMs could substitute for human annotators using the protocol introduced in Section 3.1. In each trial, we replaced one human coder at a time from Group A with an LLM, and computed Krippendorff's α for the resulting group.

To assess variability and statistical significance, we applied the paired bootstrap procedure described in Section 3.1.4. We tested

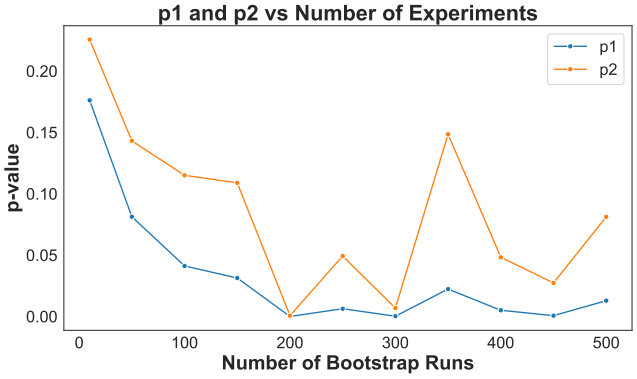

Figure 7: p-values obtained for the PolitiFact dataset for various value of B, the number of bootstrap iterations.

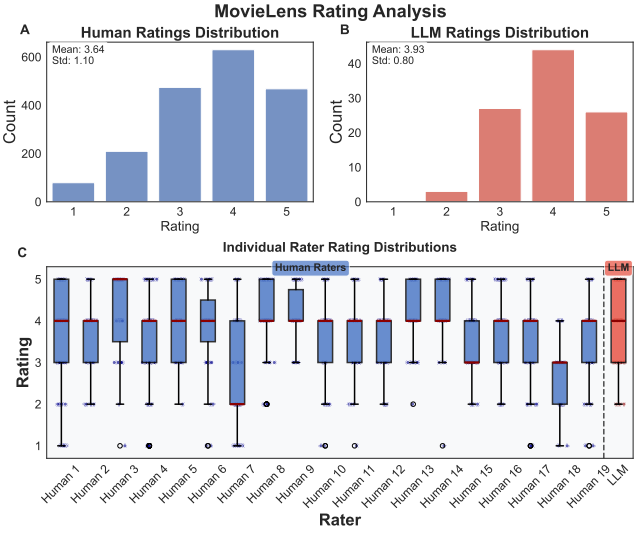

Figure 8: MovieLens ratings distribution for human and LLM.

multiple bootstrap sample sizes, with $B \in \{50, 100, \ldots, 500\}$. For each value of B, we repeated the substitution experiment **10 times**. Each repetition involved generating B bootstrap samples of 40 items per iteration, computing agreement scores, and performing the Two One-Sided t-Test (TOST) as described in Section 3.2. For each value of B, we report the *mean* and *standard deviation* over the 10 trials for the two p-values from the TOST procedure (p_1 and p_2).

We also conducted a control experiment in which human annotations were replaced with randomly generated labels, rather than LLM-generated ones. This allowed us to assess how random substitution affects inter-rater agreement and to compare its impact against that of LLM substitution.

4.4 Result

Equivalence Testing Outcomes. We assessed whether LLM-substituted annotations were statistically equivalent to human annotations across varying bootstrap sample sizes B. As shown in Figure 6

Table 2: TOST result ($\mu \pm \sigma$) for two datasets experiment for $B = 300$

Dataset	Margin	Human α	LLM α	Random α	LLM p_1	LLM p_2	Random p_1	Random p_2
MovieLens 100K	0.025 ± 0.014	0.199 ± 0.001	0.199 ± 0.001	0.164 ± 0.002	0.002	0.004	0.505	<0.001
CrowdsourcingTruthfulness–PolitiFact	0.034 ± 0.030	0.098 ± 0.004	0.100 ± 0.004	0.090 ± 0.004	0.047	0.155	0.092	<0.001

and Figure 7, the p-values exhibit different behaviors across datasets: on **MovieLens 100K**, p-values tend to decrease with larger B, indicating increased stability; on **PolitiFact**, no consistent pattern emerges.

Table 2 summarizes the final results at $B = 300$, reporting the mean and standard deviation across 10 runs for the equivalence margin, Krippendorff's α for the human group, the LLM-substituted group, and the randomly substituted group, along with the corresponding TOST p-values (p_1 and p_2) for both LLM and random substitutions. The LLM passed the equivalence test on the MovieLens 100K dataset but failed on PolitiFact—despite nearly identical α scores between the LLM- and human-only groups. This highlights how small margins and higher variability, shown in the per-rater results discussed subsequently, can lead to non-equivalence conclusions. In contrast, substituting annotations with random labels consistently produced significantly lower α scores and failed the equivalence test in both datasets, confirming that LLM-generated annotations are substantially more aligned with human judgment than random labels.

Annotation Distribution Comparisons. and Figure 8 show the rating distributions across annotators. LLM ratings align more closely with human annotations in MovieLens 100K, while notable differences are observed in PolitiFact. These distributional patterns mirror the statistical equivalence findings.

Agreement Change per Rater. Finally, we examined how Krippendorff's α changed when each individual human coder was replaced with the LLM (Figure 10, 11). For the PolitiFact dataset, changes ranged from −2.3% to 2.5%, and for MovieLens 100K, from −1.5% to 1.7%. In both datasets, some raters showed minimal change (as low as 0.1%), suggesting that the LLM closely aligned with certain individuals. The variability in the per rater change in Krippendorff's α suggests that while substitution effects are minimal on average, individual rater alignment with the LLM may vary. Additionally, the smaller variability in the changes of the per rater Krippendorf α on the MovieLens 100k dataset dataset compared to the PolitiFact dataset also indicates a closer LLM alignment to the raters for the movie rating task compared to rating the truthfulness of a piece of information.

5 Discussion

Our results demonstrate that the LLM's ability to substitute for human annotators is highly task-dependent. On the MovieLens dataset, the model passed our equivalence test, whereas on the PolitiFact dataset it failed. This divergence reflects fundamental differences between the two annotation domains: preference-based rating versus fact-checking.

For MovieLens, it's not surprising that LLM ratings align with human ratings. As shown in Figure 8, the model's scores follow the general patterns of human raters. This likely comes from its

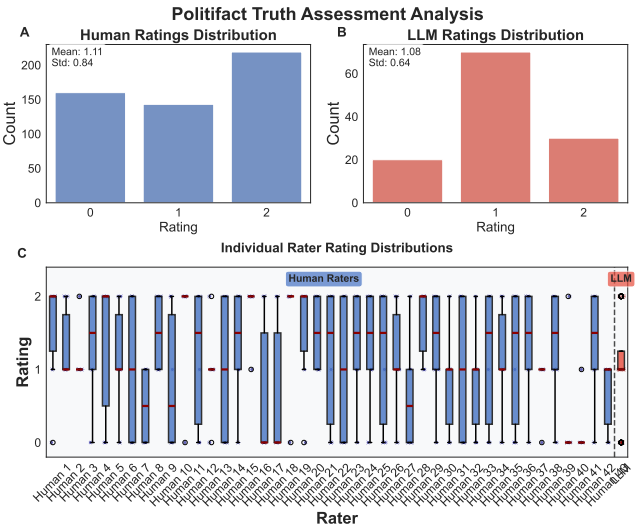

Figure 9: PolitiFact ratings distribution for human and LLM.

pretraining on large amounts of movie-related text [13], which helps it learn common rating habits and genre cues. In other words, judging movie enjoyment mostly means recognizing shared cultural opinions and repeating them. While there is some subjectivity, the range of reasonable answers is limited and well covered in the training data. That's why replacing human ratings with LLM ratings hardly changes Krippendorff's α in Figure 10.

In the PolitiFact dataset, replacing human ratings with LLM ratings lowers reliability across raters (see Figure 11). Fact-checking is harder than movie rating because it requires domain knowledge, evidence usage, and political framing awareness. Human annotators bring in their own beliefs, expertise, and even mood, which creates variability [18, 25, 45]. Different cues like familiarity, political views, or emotional reactions influence people in different ways [19, 23]. The LLM, however, takes a narrower and cautious approach: it avoids strong labels like "True" or "False" and stays near the middle categories Figure 9. This mismatch with human judgment patterns explains why inter-rater reliability drops more in this case.

The alpha-change analysis (Figures 10 and 11) shows a clear contrast. In MovieLens, swapping some human raters with the LLM has a minimal impact on reliability, indicating that the model can effectively mimic certain annotators. In PolitiFact, though, replacements cause big drops in agreement for some raters. This suggests LLMs still have trouble replicating the unique, knowledge-based reasoning humans use in complex or disputed topics.

Taken together, these findings underscore two key points. First, our evaluation method captures meaningful differences across tasks: it is mathematically rigorous yet sensitive to domain characteristics.

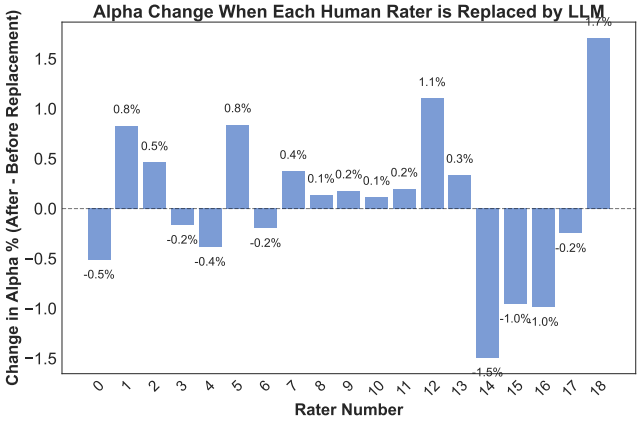

Figure 10: The change in Krippendorf's α before and after replacing for Movielens 100K dataset.

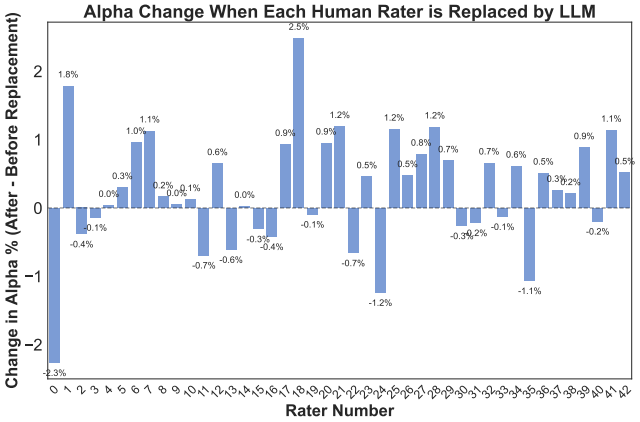

Figure 11: The change in Krippendorf's α before and after replacing for PolitiFact dataset.

Second, LLM substitution is more viable for preference-oriented annotations than for knowledge-intensive or adversarial tasks. These results highlight the importance of aligning LLM-based annotation strategies with the epistemic demands of the domain.

Guidelines. Practitioners should apply our evaluation methodology as follows. For existing datasets, replicate the workflow described in Section 4. If the dataset does not meet our minimum criteria (four annotators, at least 80 items), a small-scale annotation effort should be undertaken. The same approach applies to new tasks: collect modest and diverse annotations, compute reliability measures, and test LLM substitution feasibility before scaling.

Interpreting the Results. When an LLM passes the equivalence test, it can be considered a candidate substitute for the replaced annotator. If that annotator is a gold-standard rater, the LLM may then be deployed to expand annotations at scale. If substitution fails, human annotation remains indispensable, though alternative models or prompting strategies could be tested. Importantly, our alpha-change framework also enables more granular exploration: identifying which human rater is most closely approximated, comparing across LLMs, or diagnosing systematic biases in annotation behavior. This level of analysis extends beyond a simple pass/fail judgment, offering a roadmap for responsibly integrating LLMs into annotation pipelines.

6 Conclusion

Our experimental results demonstrate that the proposed evaluation method can effectively assess whether an LLM approximates human judgment in specific text annotation tasks, using only a small number of annotated items. We applied the method to two datasets—MovieLens 100K ($p = 0.004$) and PolitiFact ($p = 0.155$)—with differing outcomes. The LLM passed the equivalence test on MovieLens 100K but not on PolitiFact. These results highlight that performance is not consistent across tasks and depends heavily on the nature of the annotation task.

This method provides a practical way to detect differences in annotation behavior between humans and LLMs. It also offers an opportunity to evaluate a small set of annotations before deciding whether to use an LLM for large-scale annotations.

Limitations and Future Work. The performance of the LLM is likely to vary depending on the specific model and the prompts used. Future work should investigate how various LLM architectures and prompt strategies impact the results, facilitating a more comprehensive evaluation of LLM-human alignment across diverse annotation contexts.

Acknowledgments

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005), and funded fully by the Australian Government through the Australian Research Council. This work was conducted on the unceded lands of the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation. We pay our respect to Ancestors and Elders, past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their connections to land, sea, sky, and community.

References

[1] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. LLMs can be Fooled into Labelling a Document as Relevant: best café near me; this paper is perfectly relevant. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 32–41.

[2] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Information processing & management* 48, 6 (2012), 1053–1066.

[3] Miriam E Armstrong, McKenna K Tornblad, and Keith S Jones. 2020. The accuracy of interrater reliability estimates found using a subset of the total data sample: A bootstrap analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 64. SAGE Publications Sage CA: Los Angeles, CA, 1377–1382.

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

[5] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 667–674.

[6] Krisztian Balog, Don Metzler, and Zhen Qin. 2025. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. 3865–3875.

[7] Berndt Brehmer, Roger Hagafors, and Roger Johansson. 1980. Cognitive skills in judgment: Subjects' ability to use information about weights, function forms, and organizing principles. *Organizational Behavior and Human Performance* 26, 3 (1980), 373–385.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.

[9] Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. 2025. Crowdsourcing or AI Sourcing? *Commun. ACM* 68, 4 (2025), 24–27.

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

[11] Richard Craggs and M Wood. 2004. A two dimensional annotation scheme for emotion in dialogue. In *Proceedings of AAAI spring symposium: exploring attitude and affect in text*, Vol. 102.

[12] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18, 4 (2005), 407–422.

[13] Dario Di Palma, Felice Antonio Merra, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do LLMs Memorize Recommendation Datasets? A Preliminary Study on MovieLens-1M. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. 2582–2586.

[14] Laura Dietz, Oleg Zendel, Peter Bailey, Charles Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proc. ICTIR*.

[15] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1107–1128.

[16] Hillel J Einhorn. 1971. Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational behavior and human performance* 6, 1 (1971), 1–27.

[17] Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 39–50.

[18] Daniel T Gilbert, Romin W Tafarodi, and Patrick S Malone. 1993. You can't not believe everything you read. *Journal of personality and social psychology* 65, 2 (1993), 221.

[19] Kenneth R Hammond. 1955. Probabilistic functioning and the clinical method. *Psychological review* 62, 4 (1955), 255.

[20] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4, Article 19 (Dec. 2015), 19:1–19:19 pages.

[21] Jiaman He, Zikang Leng, Dana McKay, Johanne R Trippas, and Damiano Spina. 2025. Characterising Topic Familiarity and Query Specificity Using Eye-Tracking Data. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2602–2606.

[22] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 5549–5581.

[23] Paul J Hoffman. 1960. The paramorphic representation of clinical judgment. *Psychological bulletin* 57, 2 (1960), 116.

[24] 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. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Karin Verspoor, Kevin Bretonnel Cohen, Mark Dredze, Emilio Ferrara, Jonathan May, Robert Munro, Cecile Paris, and Byron Wallace (Eds.). Association for Computational Linguistics, Online.

[25] Klaus Bruhn Jensen, Robert T Craig, Jefferson D Pooley, and Eric W Rothenbuhler. 2016. *The International Encyclopedia of Communication Theory and Philosophy, 4 Volume Set*. John Wiley & Sons.

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

[27] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

[28] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology* (4th ed.). SAGE Publications, Thousand Oaks, CA.

[29] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement?. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 805–814.

[30] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement?. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 805–814.

[31] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3214–3252.

[32] Sean MacAvaney and Luca Soldaini. 2023. One-shot labeling for automatic relevance estimation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2230–2235.

[33] Angel Felipe Magnossão de Paula, J Shane Culpepper, Alistair Moffat, Sachin Pathiyan Cherumanal, Falk Scholer, and Johanne Trippas. 2025. The Effects of Demographic Instructions on LLM Personas. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3045–3049.

[34] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14867–14875.

[35] Stefano Mizzaro. 1997. Relevance: The whole history. *Journal of the American society for information science* 48, 9 (1997), 810–832.

[36] Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 18511–18526.

[37] Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9048–9062.

[38] Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access* 12 (2024), 71876–71900.

[39] Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics* 7 (2019), 677–694.

[40] Jiaxin Pei and David Jurgens. 2023. When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, Jakob Prange and Annemarie Friedrich (Eds.). Association for Computational Linguistics, 252–265.

[41] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.

[42] James L. Rogers, Kenneth I. Howard, and John T. Vessey. 1993. Using Significance Tests to Evaluate Equivalence Between Two Experimental Groups. *Psychological Bulletin* 113, 3 (1993), 553–565.

[43] Kevin Roitero, David La Barbera, Michael Soprano, Gianluca Demartini, Stefano Mizzaro, and Tetsuya Sakai. 2023. How many crowd workers do I need? On statistical power when crowdsourcing relevance judgments. *ACM Transactions on Information Systems* 42, 1 (2023), 1–26.

[44] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Information Processing & Management* 58, 6 (2021), 102688.

[45] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[46] Tetsuya Sakai. 2025. My System Is As Effective As Yours: Reproducibility, Sustainability, and More. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. 3943–3953.

[47] S. Salvador and P. Chan. 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE International Conference on Tools with Artificial Intelligence*. 576–584.

[48] Julian A Schnabel, Johanne R Trippas, Falk Scholer, and Danula Hettiachchi. 2025. Multi-stage large language model pipelines can outperform gpt-4o in relevance assessment. In *Companion Proceedings of the ACM on Web Conference 2025*. 1288–1292.

[49] Marco Siino, Ilenia Tinnirello, Marco La Cascia, et al. 2025. From Foundations to GPT in Text Classification: A Comprehensive Survey on Current Approaches and Future Trends. *Foundations and Trends® in Information Retrieval* 19, 5 (2025), 557–711.
[50] Steven M Snapinn. 2000. Noninferiority Trials. *Trials* 1, 1 (July 2000).
[51] T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay. 2021. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics. arXiv:2105.11910 [cs.CL] https://arxiv.org/abs/2105.11910
[52] Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability* 57, 1 (1993), 1–436.
[53] Alan M Turing. 2009. *Computing machinery and intelligence*. Springer.
[54] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 422–426.
[55] Oleg Zendel, J Shane Culpepper, Falk Scholer, and Paul Thomas. 2024. Enhancing human annotation: Leveraging large language models and efficient batch processing. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 340–345.
[56] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, Mexico City, Mexico, 3881–3906.