ARE GROUND TRUTH LABELS REPRODUCIBLE? AN EMPIRICAL STUDY

Ka Wong[1], Praveen Paritosh[1], and Kurt Bollacker[2]

[1]Google Research - {kawong,pkp}@google.com
[2]Long Now Foundation - kurt@longnow.org

ABSTRACT

Standard evaluation techniques in Machine Learning (ML) and the corresponding statistical inference typically do not typically explicitly consider ground truth labels as a source of uncertainty. They are assumed to be reproducible. This is not necessarily true. To demonstrate that, we investigate the reliability of nine test sets used in Northcutt et al. (2021b). They contain items sampled from highly cited benchmarks that are particularly challenging. Indeed, they are challenging for human to label as well. By replicating Northcutt et al.'s experiments, we find the ground truth labels, obtained via simple majority votes of five, to have zero reliability in three of the experiments. They are statistically random. The cause of irreproducibility is excessive rater disagreement, as evidenced in the zero inter-rater reliability (IRR). Contrary to popular belief, majority voting fails to boost the signal in these instances. To explore a solution, we conducted a smaller pilot using raters with high qualifications and found a significant improvement in reliability across the board. This shows *high quality data collection is still paramount, and cannot be supplanted by aggregation*. We urge researchers, reviewers, and publishers (such as for conferences and journals) to encourage the measurement and reporting of reproducibility in the ground truth data used in ML evaluation. Towards this end, we publish and release all the replications and associated data to aid assessment of reproducibility of this work.

1 INTRODUCTION

Human labeled benchmarks are the primary empirical basis for comparing models and assessing the progress on most ML tasks. Schaekermann et al. (2020) estimated over 60% of the papers in NeurIPS 2020 utilized human labeled ground truth data. In this paper, we investigate the reproducibility of the ground truth generating process that is often used to produce these benchmarks.

The process of generating ground truth data consists of aggregating human annotations, typically recruited from crowdsourcing platforms (Geva et al., 2019; Sabou et al., 2014). Its relatively low costs and scalability led to the creation of many important datasets such as the ImageNet (Deng et al., 2009) and FreeBase (Bollacker et al., 2008). Such advantages come at a cost: data quality. From the onset, researchers have worried about the quality of this data, prompting early investigations into understanding how labeling redundancy and aggregation can help mitigate quality concerns (Snow et al., 2008; Sheng et al., 2008; Nowak and Rüger, 2010). These studies conclude that the aggregate performance of a small number of non-expert annotations roughly match that of expert annotators. Slowly this "replicate and aggregate" strategy of generating labeled data has become synonymous with *ground truth* data collection. It is now the de facto data collection method in ML and is the reference for model evaluation.

The recent rise of ML has seen a proliferation of increasingly more subjective tasks such as affective computing (Cowen et al., 2021), hate speech detection (Leonardelli et al., 2021), and natural language inference (Nie et al., 2020). In these tasks, subjective human judgements play an central role. At the same time, the rapid pace of innovation in ML means that data are collected under lots of different experimental conditions, e.g. with various degrees of rater quality control (Nowak and Rüger, 2010).

Together, this means the random factors in the labeling process such as the sampling of raters and their spontaneous decision-making can have an outsize impact on the data collection. With such uncertainty in the labeling process, can we rely on aggregation to produce *reliable* ground truth labels across a spectrum of tasks, by using just a handful of labels per item?

Ground truth reliability rarely enters our mind set. While reliability of the unaggregated data can be readily computed from redundant labels, ground truth labels are not commonly replicated. Unfortunately, comparisons of aggregation techniques only focus on *accuracy* and fail to consider reliability as a comparison criterion. This lack of awareness is further evidenced in the fact that many papers are concerned with ways to cope with imperfect data in *training* (Rolnick et al., 2017), they often assume the *test* sets are pristine. In short, there is a gap in the literature on the reliability quantification of ground truth labels. This is an important question as it affects our confidence in the evaluation. These are the motivations of this paper. We study this problem using a sample of nine verification datasets published in Northcutt et al. (2021b). We measure the reliability of ground truth labels via replication and show they can be highly unreliable. In summary, the contributions of this work are:

1. We replicated nine verification datasets to empirically measure the reliability of their ground truth labels.
2. We show that inter-rater reliability is a strong driver behind ground truth reliability and that aggregation's affect is limited.
3. We show that high ground truth reliability is achievable through better data collection practices, e.g. using raters with higher qualification.

2 Related Work

In this section we summarize work addressing some of the sources of irreproducibility in ML research.

2.1 Sources of Uncertainty in Model Evaluation

The degree of certainty in model evaluation determines our ability to differentiate models based on their genuine capabilities (Card et al., 2020). A typical model evaluation consists of a sample of labeled items and predicted scores. The sampling of the items, the labeling process, and the generation of scores are all potential sources of uncertainty. However, in the literature, many papers focus on the items and the model, and fail to consider label reliability. In *Accounting for Variance in Machine Learning Benchmarks*, Bouthillier et al. (2021) discuss variance in evaluation due to items and model architecture, but omits the fact that labels are not perfectly reliable. Dror et al. (2018) surveys a number of methods for statistical significance testing, but the methods only account for the sampling of items and do not explicitly consider the reliability of labels. There are exceptions. In Card et al. (2020), the authors study the statistical power of various Natural Language Processing (NLP) benchmarks, and they use a random effects model to capture the variability of the labels. In *The Benchmark Lottery* (Dehghani et al., 2021), the authors acknowledge the selection of rater pool in evaluation may impact the model performance. Finally, Bowman and Dahl (2021) stress the importance of reliable test data in benchmarking. They argue lack of data reliability will ”compromise the validity of the benchmark.” Yet, this is far from standard practice; on the contrary, most widely used ML benchmarks are of unknown reliability.

2.2 Factors of Ground Truth Reliability

Ground truth labels are the result of **aggregation** over **a number** of **raters**.[1] Naturally, as we seek to understand the drivers behind ground truth reliability, we should consider 1) the aggregation method, 2) rating redundancy, and 3) rater disagreement. In our replication experiments, we focus exclusively on 3) rater disagreement as a first demonstration. We do not alter the aggregation method or rating redundancy used in the original study. We take these design parameters as fixed in order to isolate and highlight the impact of rater disagreement on ground truth reliability. Nevertheless, we discuss all 3 factors below.

[1] The number of workers can vary between items, but for simplicity we assume it is fixed.

2.3 Rater Disagreement

Rater disagreement is a central concern in any human labeling experiment. Many factors affect disagreement, such as qualifications of the raters, their personal and cultural background (Prabhakaran et al., 2021), guideline insufficiency (Bowman and Dahl, 2021), and most importantly, the inherent task ambiguity. In NLP, Nie et al. (2020), Leonardelli et al. (2021), and Pavlick and Kwiatkowski (2019) found abundant disagreement even after elaborate rater quality control. This speaks to the genuine inherent ambiguity of the task. In image recognition, 20% of images in ImageNet were found by expert labelers to contain more than one prominent subject (Shankar et al., 2020). Researchers are beginning to provide multi-class labels as a solution (Beyer et al., 2020; Rodriguez et al., 2021; Shankar et al., 2020).

In scientific disciplines such as linguistics and psychometrics that routinely rely on human annotated data, inter-rater reliability is used to quantify rater disagreement and to ensure the reproducibility (Paritosh, 2012). Recommendations such as Landis and Koch (1977) are often used as publication criteria.

2.4 Aggregation Techniques

Also known as *truth inference* (Zheng et al., 2017; Difallah and Checco, 2021), label aggregation is used to counter labeling quality issues to recover the true labels. We focus on majority voting in this paper (Sheng et al., 2008) to be consistent with the Northcutt et al. (2021b) paper. The mean (Snow et al., 2008) is another popular option.

Many different aggregation techniques have been proposed: Difallah and Checco (2021) gives a recent review of techniques for different task types (e.g. rating scales, image vs text). Other methods have been proposed to make the best inference based on different experimental assumptions (e.g. error prevalence, independence of error with respect to raters and items, rater weighting schemes, different priors, and sparsity of data). A notable technique worth mentioning is Dawid and Skene (1979)'s EM-style iterative algorithm that considers individual raters' performance, using the weighted mean in each M-step as a stand-in for ground truth. Also worth mentioning is its Bayesian variant Raykar et al. (2009).

With such a proliferation of different techniques, choosing the appropriate one to use is an open-ended question. Sheshadri and Lease (2013) compare seven popular methods and conclude none of the aggregation method is consistently the best across experiments. Zheng et al. (2017) come to a similar conclusion comparing 17 methods. Both papers note that majority voting provides a reasonable, easy-to-implement baseline, but more sophisticated methods can potentially provide additional benefits.

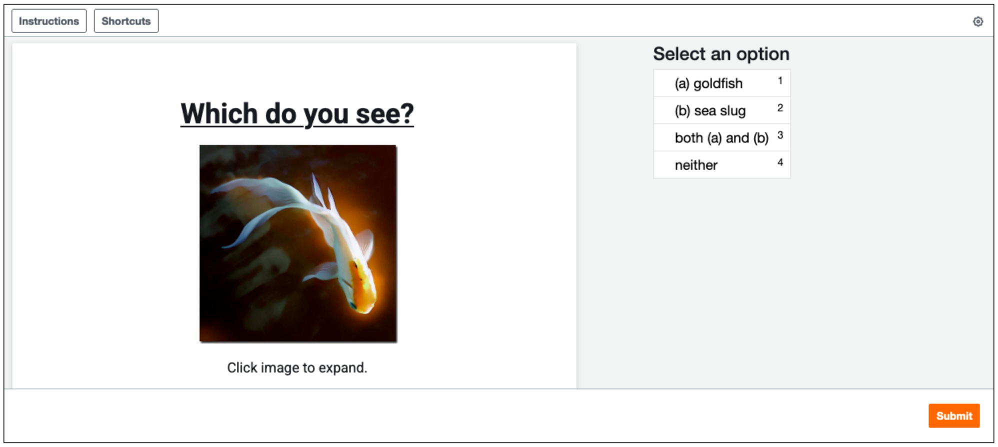
We note that these comparison studies only focus on *accuracy* of the aggregation methods, which is often described circularly Riezler (2014), and fail to consider their reliability. Hence we do not know how these methods differ in terms reliability in their final outputs.

2.5 Rating Redundancy

Past large-scale demonstrations of *Wisdom of Crowds* tended to employ high rating redundancies. Galton (1907)'s first demonstration involved 787 farmers at a county fair. Simoiu et al. (2019) uses 1000 ratings per question to test the validity of Wisdom of Crowds.

There is a reason for that. In the case of the mean, the reliability of mean labels strictly increases as a function of the rating redundancy (Ebel, 1951). The functional form is given by the Spearman-Brown formula (de Vet et al., 2017) and the formula is routinely used by behavioral scientists in experimental design. Based on this, we have evidence that reliability of ground truth labels depends on the rating redundancy, regardless of the aggregation methods.

Unfortunately, rating redundancy in crowdsourcing experiments is typically low. Card et al. (2020) conducted a meta-analysis of EMNLP papers that utilize human evaluation for model comparison. They find 57% of experiments collected three annotations per item. Such low rating redundancies is the norm rather than the exception. Rating redundancy is an extra safeguard against labeling uncertainty, but the current levels may not be sufficient.


Which do you see?

Click image to expand.

Figure 1: A screenshot of the ImageNet verification tasks. One label is taken from ImageNet, one is a predicted label produced from confident learning (Northcutt et al., 2021a).

3 Method

We describe the nine original Northcutt et al. (2021b) datasets,[2] how we replicated them, and our methodology for measuring the datasets' inter-rater reliability and ground truth reliability.

3.1 Original Verification Datasets

We build up on the work by Northcutt et al. (2021b), which investigated 10 highly cited ML benchmarks test sets from a broad set of tasks. This work went above and beyond the state-of-the-art in ground truth data generation, and publicly released all their data and methodology. They found pervasive errors in these benchmarks: an estimated average of at least 3.3% errors across the 10 datasets, e.g., 6% of the ImageNet test set. In this section we summarize their approach to identifying errors in the benchmarks.

They first used confident learning (Northcutt et al., 2021a) to identify test set items that have a *highest likelihood of having an incorrect label.* In other words, this is not a random sample of the original benchmarks, also likely the hardest to label subset for human raters as well. This makes this dataset really valuable for reproducibility research. These items were then sent to Mechanical Turk (MTurk) to be validated in a human evaluation experiment. Each item was shown to 5 MTurk raters along with two different labels: one from the benchmarks and one predicted from confident learning. The raters selected whether they saw the (1) benchmark label, (2) the predicted label, (3) both labels, or (4) neither in the example. A screenshot of the task is shown in Figure 1. This is not a typical design as it involves two labels (one machine generated candidate, and the other human generated from the original benchmarks) instead of one. That is because the authors were trying to identify errors in the benchmarks and simultaneously correct them using the predicted labels. The authors declare the original benchmark label correct if three or more workers choose option (1). This amounts to a simple majority of 5. We are interested in studying this verification process, hence we will only concern ourselves with option (1), and treat other options as not (1). Specifically, after we collect the data, we transform it into a binary verification task of confirming the benchmark labels. There are additional details regarding the benchmarks that we will not reproduce here. We refer interested readers to Northcutt et al. (2021b) for more details.

The authors released all their collected labels and invited readers to browse and contribute further improvements. We use them as a starting point for this investigation. Not only do these datasets present a spectrum of tasks, but the authors go above and beyond the state-of-the-art in building ground truth data, e.g., Krishna et al. (2016), as well as in the methodological reproducibility of

[2] The original study had ten datasets, and we had some technical difficulties in replicating one of them, the Audioset dataset from Gemmeke et al. (2017).

making data and process publicly available. We are grateful for the generous data sharing by the authors and follow and encourage others to adopt this open practice.

3.2 Replications

We re-labeled nine of the ten aforementioned datasets using MTurk. To ensure we replicated the experiments faithfully, we communicated with the authors to learn about additional details not reported in their paper. We followed each step carefully, including rating templates, rater selection criteria, and rater compensation. As in the original study we assigned five MTurk raters to each item. Our goal in replication is to measure the reliability of the ground truth generation process, so we try to mimic the original conditions to our best ability. All the datasets will be made available on GitHub (not yet ready at the time of publication): `https://github.com/google-research-datasets/ground-truth-replications`.

3.2.1 Challenges in replication

Some of the tasks in the original study presented raters with canonical examples of each label. The specific examples used in the study were not published, so we experimented with a few heuristics based on recommendations from the authors: a) choosing the highest predicted class probability for that sample, b) choosing the samples that had the lowest class entropy, c) choosing the examples from the original benchmarks as the canonical source. We noticed variation in results depending on the heuristics for choosing examples to show. Thus, in this paper we decided to report the rater data without examples, for the sake of consistency across all the datasets. We report on 18 replications (nine regular, nine masters) without examples. We publish all the 30 replications we tried (previous plus three ImageNet replications with examples, and nine masters without examples) in the associated data release. Despite our best effort, there are unavoidable differences. For example, the worker pools may have shifted in the year since.

3.3 Raters With Higher Qualifications

In addition to replicating the original datasets, we ran a small pilot to understand the impact of using more qualified raters on data reliability. We elected “masters” raters on the MTurk platform, and among them only chose those with 90% acceptance rate and a minimum experience of 100K HITs. Unlike our replications in the last section, we assigned 10 MTurk workers to each item (five in each of the two replications), in order to measure the reliability of the resulting ground truth labels (see Section 3.5 below). This is merely a proof-of-concept, so we only collected masters labels on 100 items in each benchmark dataset that are randomly sampled.

3.4 Inter-Rater Reliability

In a labeling experiment, the extent of rater agreement is quantified using *inter-rater reliability* (IRR). Reliability is preferred over %-agreement because the latter is sensitive to class-imbalance and can lead to overly optimistic results (Chicco and Jurman, 2020). IRR is the canonical measure of rater disagreement in psychometrics (Brennan and Hays, 1992) and linguistics (Artstein and Poesio, 2008) and is use as a primary safeguard against irreproducibility of human labeled data (Paritosh, 2012).

Many reliability coefficients exist to accommodate different experimental conditions and assumptions. Common measures include Cohen’s (1960) *kappa*, and intraclass correlation (McGraw and Wong, 1996), and Krippendorff’s *alpha* (Krippendorff, 2011). Our coefficient of choice is Krippendorff’s *alpha* as it can accommodate multi-rater, binary-rating experiments. Krippendorff’s *alpha* requires the raters be *interchangeable*, which we feel is justified in the original experiments with a large number of MTurk raters. We calculate Krippendorff’s *alpha* as

$$\alpha = 1 - \frac{1 - \text{Observed Agreement}}{1 - \text{Chance Agreement}},$$

where observed agreement is the proportion of times two raters agree, and chance agreement is expected agreement by chance. Reliability ranges from 0 to 1. IRR $= 1$ indicates perfect agreement between raters. IRR $= 0$ indicates there is no observable agreement between raters beyond chance agreement. Different minimum acceptable reliability values have been suggested, e.g. Carletta (1996)

Benchmark Label	**Original Ground Truth**	**Replicated Ground Truth**
Apple	TRUE	TRUE
Apple	FALSE	TRUE
Orange	TRUE	FALSE
Orange	FALSE	TRUE
Apple	FALSE	FALSE
Orange	TRUE	FALSE

Table 1: Schematics illustrating the evaluation of benchmark labels. The Benchmark column is original labels to be verified by MTurk raters. Their collective judgements are aggregated into ground truth labels via majority vote of five. Original and Replicated Ground Truth columns represent the original Northcutt majority votes and our replicated results respectively. Reliability of ground truth labels is computed as the Krippendorff's α between them.

suggest IRR $= 0.67$ for computation linguistics and Landis and Koch (1977) suggest IRR $= 0.6$ for medical sciences.

3.5 Ground Truth Reliability

To compute IRR, a minimum of two ratings per item is required. In Northcutt's verification datasets, each item is rated by five raters, hence IRR can be computed readily from the published data. For *ground truth reliability*, however, additional data is required. This is because each verification dataset only produces one vector of ground truth labels, hence we have to replicate the experiments in order to obtain another instantiation. There is an exception. If the aggregation method is the mean, then one can use the Spearman-Brown formula (de Vet et al., 2017) to analytically calculate the reliability of mean ratings based on the IRR. For majority voting, the authors are not aware of any closed-form solution and have to resort to empirical replication.

Similarly as before, we used Krippendorff's *alpha* to calculate the reliability of ground truth labels. To justify this choice, we checked the marginal distributions to ensure there are no significant distributional shifts between the two sets of experiments. We illustrate the computation of α in Table 1. The first column is the benchmark labels being validated. After collecting the labels, we binarize them into TRUE/FALSE based on whether they agree with the benchmark label as described in the previous section. Original Ground Truth represents the 5-rater simple majority votes from Northcutt's experiments. Replicated Ground Truth represents their counterparts from our own replications. Reliability of the ground truth labels is taken to be Krippendorff's *alpha* between these two ground truth columns.

4 Results

We present results from our replication experiments in this section. We show that the IRRs in both the original and replicated verification datasts are low. Some are effectively zero (random), leading to equally low ground truth reliability. The vision tasks are found to have higher IRR and benefit more from majority voting. We illustrate that IRR is a strong driver behind ground truth reliability. Lastly, we show a proof-of-concept demonstration that using more qualified raters can lead to large gains in reliability.

4.1 Inter-Rater Reliability

The IRR values of the original and replicated verification datasets are shown in Table 2. The IRR values are generally quite low. In particular, the three NLP tasks are not statistically significantly different from zero. In other words, the data are effectively indistinguishable from a random shuffle. Perhaps this is because the Amazon and IMDB tasks are sentiment analysis, so they are expected to be more subjective. However, the excessive rater disagreement in these tasks suggests they may be under-defined or the experiments were otherwise not well executed.

On the other hand, vision tasks are generally thought to be more objective. Yet, they are exhibiting low IRR ranging from 0.178 to 0.474, which is still below the 0.6 cutoff recommended by Landis

Benchmark	**Modality**	**Size**	**IRR (orig.)**	**IRR (repl.)**	**Delta**
20news	text	93	**-0.005***	**-0.025***	-0.020
Amazon	text	1,000	**0.033***	**0.013***	-0.020
IMDB	text	1,310	**0.144***	**0.077***	-0.067
CIFAR-10	image	275	0.178	0.164	-0.014
QuickDraw	image	2,500	0.245	0.231	-0.014
CIFAR-100	image	2,235	0.290	0.244	-0.046
MNIST	image	100	0.377	0.348	-0.029
Caltech-256	image	400	0.423	0.375	-0.048
ImageNet	image	5,440	0.474	0.224	**-0.250**

*not statistically significant at 95% confidence

Table 2: IRR of the nine original verification datasets and their replicated counterparts. Delta = replicated IRR - original IRR. The bolded values are not statistically significant. All the replicated IRRs are smaller than the originals, despite our best effort. All the deltas are relatively small, with the exception of the ImageNet.

and Koch (1977). The low reliability may be partially explained by the inherent ambiguity in the images. For example, the CIFAR-10, QuickDraw, and CIFAR-100 have lowest IRR among vision tasks. CIFAR images are tiny, and QuickDraw images are rough hand-drawn sketches. These types of images can be prone to rater disagreement due to inherent ambiguity. This shows that the subset identified by Northcutt et al. (2021b) are not only difficult for machines but also difficult for human raters to consistently label. Far from a random sample of the original benchmarks, the reliability results derived from this sample do not reflect the reliability of the original benchmarks. However, as this represents challenging subsets of these benchmarks, any claims of machine performance on this subset are hard to quantify due to the lack of labelling reproducibility. Simply put, this subset identified by Northcutt et al. (2021b) is both challenging for machines to predict and for humans to verify. How do we deal with this last mile in ML evaluation?

In the replications, the replicated IRR tracks the original values fairly well, as evidenced in the relatively small deltas. However, there are two major concerns. First, all the replicated IRRs are consistently smaller than the originals. Secondly, the IRR of the ImageNet replication is significantly lower than the original. This means our replications were not done in an identical fashion as the original experiments, despite our attempt to follow the exact instructions from the original authors. We acknowledge the shortcomings in our replication process in

5 Challenges in Replication

, but the results are corroborated with IRR. These problems speak to the challenges in running replication experiments in general, despite the original authors' generous sharing of data and methodology. We acknowledge this is a weakness in our study and will continue to investigate it.

5.1 Ground Truth Reliability

In the previous section, we looked at the reliability of the rater labels. In this section, we look at the reliability of the ground truth data, i.e., the ground truth labels obtained by aggregating the five votes using majority. The ground truth reliabilities in the nine verification datasets are shown in Table 3. Observed agreement and chance agreement are also shown to help make the calculation of reliability more transparent. Nie et al. (2020) use 1 - %-agreement between replications of ground truth labels to measure their uncertainty and call it "change rate." Despite the possibility of being mis-interpreted, we include %-agreement for greater transparency. To ensure proper interpretation, we include the chance agreement as a baseline. The chance agreement is calculated according to Krippendorff (2011). It is implicitly used in the calculation of IRR, but we make it explicit here for ground truth labels.

As foreshadowed, the three NLP tasks with low IRR turned out to have low, non-significant ground truth reliability. This implies the "true label" of each item in these verification sets is only known up to its marginal distribution, and is effectively indistinguishable from a random baseline. This is bad

Benchmark	**Observed Agreement (%)**	**Chance Agreement (%)**	**Reliability** (α)
20news	76.3	74.2	**0.075***
Amazon	64.8	62.0	**0.074***
IMDB	59.0	55.9	**0.071***
CIFAR-10	82.5	71.7	0.384
QuickDraw	76.7	61.9	0.387
CIFAR-100	79.9	64.8	0.428
MNIST	86.0	70.4	0.526
Caltech-256	87.2	69.3	0.582
ImageNet	71.8	50.2	0.435

*not statistically significant at 95% confidence

Table 3: Reliability of ground truth labels in nine verification datasets. The bolded values are not statistically significant. Observed agreement and chance agreement (baseline) are shown to help make the calculation of reliability more transparent, as α is equal to 1 - (1 - observed agreement) / (1 - chance agreement). Note the chance agreement figures are high due to class imbalance.

news for the model builders. If they rely on the these labels to try to intuit their model's strengths and weaknesses, and to use this information to make improvements to the model, it will no different than shots in the dark.

As promising as aggregation is, majority-voting in the NLP experiments fails to result in any material improvement in reliability. In the vision tasks, it is more effective. The ground truth reliability in vision is higher, ranging from 0.384 to 0.582, showing mild but consistent improvement from the IRR. However, these values are still below the recommended 0.6 cutoff.

These results together show that rater reliability is a strong driving factor of ground truth reliability. While aggregation generally can help, it simply cannot replace high quality data collection. Particularly, in experiments with excessive rater disagreement, aggregation may have a very limited effect.

5.2 Using Raters With Higher Qualifications

Many factors may have contributed to the low reliability in these experiments, and some of them may be unique to the specific experiments. For example, sentiment analysis and tiny images are likely more subjective than other experiments. While it may be possible to prescribe solutions to each individual experiment, rater quality control is generally helpful in improving reliability.

We illustrate the impact of rater quality control on reliability by re-running the same experiments using more qualified raters. Table 4 shows the new IRR and ground truth reliability resulting from using more qualified raters. There is an increase in reliability in both categories across *all* nine experiments. In particular, there are large jumps in the IRR of IMDB (+ 0.351), Amazon (+ 0.273), and Caltech-256 (+ 0.239). These also translate into large jumps in ground truth reliability correspondingly: IMDB (+ 0.308), Amazon (+ 0.570), and Caltech-256 (+ 0.174). In addition, there are large improvements in the ground truth reliability of 20news (+ 0.231), CIFAR-100 (+ 0.349), and QuickDraw (+ 0.280). Together, four of the nine experiments achieve ground truth reliability that is above the 0.6 cutoff.

We certainly do not think this is the only way to improve data reliability, nor have we tried to tailor specific improvements to each experiment. This simple exercise merely serves to illustrate that reliable ground truth is achievable, and that the key is high quality data collection, not just aggregation. We believe well-designed experiments executed by qualified raters are the key to reproducible ground truth data.

6 Discussion

We show that ground truth labels in nine experiments generated from a typical ground-truth-generating process can be highly unreliable. Using these as test labels in evaluation will render our evaluation less reproducible. We show that inter-rater reliability has a strong influence on ground truth reliability, and consequently using raters with higher qualifications helps. Given this, metrics that are agnostic

	Inter-Rater Reliability			Ground Truth Reliability		
Benchmark	**Old**	**New**	**Delta**	**Old**	**New**	**Delta**
20news	-0.005	0.140	**+ 0.145**	0.075	0.306	**+ 0.231**
Amazon	0.033	0.305	**+ 0.273**	0.074	0.644	**+ 0.570**
IMDB	0.144	0.495	**+ 0.351**	0.071	0.379	**+ 0.308**
CIFAR-10	0.178	0.303	**+ 0.125**	0.384	0.457	+ 0.073
QuickDraw	0.245	0.411	**+ 0.166**	0.387	0.668	**+ 0.280**
CIFAR-100	0.290	0.324	+ 0.080	0.428	0.777	**+ 0.349**
MNIST	0.377	0.466	+ 0.089	0.526	0.578	+ 0.052
Caltech-256	0.423	0.662	**+ 0.239**	0.582	0.756	**+ 0.174**
ImageNet	0.474	0.488	+ 0.014	0.435	0.444	+ 0.010

Table 4: New inter-rater reliability and ground truth reliability based on raters with higher qualifications are shown next to the old values. In each category, delta = new - old. All deltas are positive, meaning there is an increase in reliability in both categories across all experiments. Delta values larger than 0.1 are bolded.

to ground truth reliability may be fundamentally unfit for characterizing the quality of the data. Some of the questions that we hope that this study provokes discussion on are:

1. How do we encourage (require!) reporting the reliability of the human labeled data in the publication process?
2. How do we incentivize investment in improving reliability of widely used benchmarks, since so much rests on them?
3. How do we live with the lack of reproducibility in the ground truth? How do we do rigorous evaluation in the face of uncertainty in the ground truth data?

References

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

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. 2021. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pages 747–769.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Patricia Flatley Brennan and Bevely J Hays. 1992. Focus on psychometrics the kappa statistic for establishing interrater reliability in the secondary analysis of qualitative clinical data. *Research in nursing & health*, 15(2):153–158.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

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

Alan S Cowen, Dacher Keltner, Florian Schroff, Brendan Jou, Hartwig Adam, and Gautam Prasad. 2021. Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589(7841):251–257.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Henrica CW de Vet, Lidwine B Mokkink, David G Mosmuller, and Caroline B Terwee. 2017. Spearman–brown prophecy formula and cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85:45–49.

Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Djellel Difallah and Alessandro Checco. 2021. Aggregation techniques in crowdsourcing: Multiple choice questions and beyond. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4842–4844.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Robert L Ebel. 1951. Estimation of the reliability of ratings. *Psychometrika*, 16(4):407–424.

Francis Galton. 1907. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

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

Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. 2016. Embracing error to enable rapid crowdsourcing. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3167–3179.

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

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021a. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021b. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Praveen Paritosh. 2012. Human computation must be reproducible. In *WWW 2012, Lyon.*

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

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual international conference on machine learning*, pages 889–896.

Stefan Riezler. 2014. On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, 40(1):235–245.

Pau Rodriguez, Soumye Singhal, David Vazquez, Aaron Courville, et al. 2021. Overcoming label ambiguity with multi-label iterated learning.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. 2017. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866. Citeseer.

Mike Schaekermann, Christopher Homan, Lora Aroyo, Praveen Paritosh, Kurt Bollacker, and Chris Welty. 2020. The ai bookie—place your bets: Will machine learning outgrow human labeling? *AI Magazine*, 41(4):123–126.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2020. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Aashish Sheshadri and Matthew Lease. 2013. Square: A benchmark for research on computing crowd consensus. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1.

Camelia Simoiu, Chiraag Sumanth, Alok Mysore, and Sharad Goel. 2019. Studying the “wisdom of crowds” at scale. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 171–179.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y 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.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552.