

What’s the Meaning of Superhuman Performance in Today’s NLU?

Simone Tedeschi^{1,2}, Johan Bos³, Thierry Declerck⁴, Jan Hajič⁵,
Daniel Hershcovich⁶, Eduard H. Hovy^{7,8}, Alexander Koller⁹, Simon Krek^{10,11},
Steven Schockaert¹², Rico Sennrich^{13,14}, Ekaterina Shutova¹⁵, Roberto Navigli²

¹Babelscape ²Sapienza University of Rome ³University of Groningen
⁴German Research Center for AI (DFKI) ⁵Charles University ⁶University of Copenhagen
⁷University of Melbourne ⁸Carnegie Mellon University ⁹Saarland University
¹⁰Jožef Stefan Institute ¹¹University of Ljubljana ¹²Cardiff University
¹³University of Zurich ¹⁴University of Edinburgh ¹⁵University of Amsterdam
{tedeschi, navigli}@diag.uniroma1.it, johan.bos@rug.nl
declerck@dfki.de hajic@ufal.mff.cuni.cz dh@di.ku.dk
hovy@cmu.edu koller@coli.uni-saarland.de simon.krek@ijs.si
schockaerts1@cardiff.ac.uk sennrich@cl.uzh.ch e.shutova@uva.nl

Abstract

In the last five years, there has been a significant focus in Natural Language Processing (NLP) on developing larger Pretrained Language Models (PLMs) and introducing benchmarks such as SuperGLUE and SQuAD to measure their abilities in language understanding, reasoning, and reading comprehension. These PLMs have achieved impressive results on these benchmarks, even surpassing human performance in some cases. This has led to claims of superhuman capabilities and the provocative idea that certain tasks have been solved. In this position paper, we take a critical look at these claims and ask whether PLMs truly have superhuman abilities and what the current benchmarks are really evaluating. We show that these benchmarks have serious limitations affecting the comparison between humans and PLMs and provide recommendations for fairer and more transparent benchmarks.

1 Introduction

In recent years, research in the field of Natural Language Processing (NLP) has been driven by a frantic race to reach the top spot in popular benchmarks (Wang et al., 2018, 2019; Lai et al., 2017; Rajpurkar et al., 2018; Reddy et al., 2019). Typically the race takes the shape of a rapid cycle of parameter tuning updates by several teams, communicating their results using a shared leaderboard. Not infrequently, systems achieve better-than-human performance on several tasks (see Figure 1). Yet what does this level of performance really mean for NLP? The

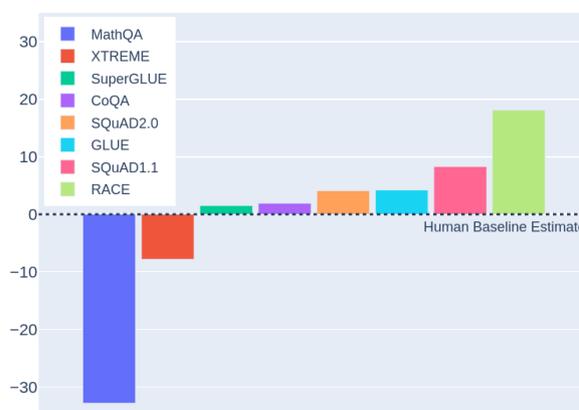


Figure 1: Difference between the scores obtained by the best-performing systems and humans in various popular NLP benchmarks. The systems outperform humans on 6 out of 8 of the reported benchmarks (best seen in color).

impressive capabilities of ChatGPT make this question even more urgent.

It is relatively easy to outperform humans with simple procedural tasks like arithmetic and extreme memory-intensive tasks involving vast amounts of data. But most tasks involving natural language typically require knowledge and inference. Do high-performing NLP algorithms really have (super)human capabilities? Or are the metrics that deliver these scores suspect?

Given the impact of claiming superhuman performance, it is important for researchers to understand exactly what is going on. As many in NLP have experienced, the false sense of accomplishment of superhuman performance often leads to an abrupt disappointment when a supposedly superb system

is applied to realistic data in a real-world situation. By propounding unrealistic claims, NLP researchers harm themselves and the field as a whole.

Some problems result from the metrics used to assess systems, which are invariably automated, and the data these metrics employ, which may be skewed in various ways. The metrics might give incomplete or biased reports of performance, or simply not apply in certain situations.

Other problems arise from the ‘boundary parameters’ that shape the task, which are usually not adequately reflected in the evaluation metric, very seldom in the kinds of automated metrics used in leaderboards. Specifically, the correctness of a task setup and its dataset instances should not be taken for granted. Also, humans and machines are often evaluated under different conditions, such as the level and type of knowledge provided to perform the task and the test items used to compute performance.

Yet other problems result from the nature of leaderboard-based evaluation. Despite the obvious benefit of driving development through competition with little human effort, these evaluations typically do not foster understanding. Teams driven by a rapid evaluation turnaround cycle in a competitive mode tend to focus more on quantitative results than on error analyses which aim at improving awareness of their problem. As currently set up, benchmarks and comparisons do not incentivize a deeper understanding of the systems’ performance, nor do they foster research geared towards producing automatic explanations: it is one thing to produce a numerical system performance score, but quite another to rate the adequacy and understandability of an explanation.

In this paper, we explore the interpretation of the superhuman performance and the utility of leaderboards, discuss how human performance is actually computed in a range of tasks, and how requirements differ for humans and automatic systems across tasks. We hope to encourage leaderboard creators to be more circumspect when setting up their challenges and provide clear ‘boundary conditions’ and descriptions of the limitations of their evaluations.

2 Popular Leaderboards Are Saturated

Leaderboard-based evaluation has become a popular practice in NLP¹ (Wang et al., 2018, 2019; Lai

¹We mainly focus on leaderboards for Natural Language Understanding (NLU) due to their commonalities, but note

et al., 2017; Rajpurkar et al., 2018; Reddy et al., 2019). The goal of these leaderboards is to encourage the development of systems capable of solving certain tasks and to measure their progress by comparing the best systems against humans. Their great success has led many researchers to focus on just the proposed tasks, resulting in a rapid saturation of the scores which, in many tasks, are equal to or greater than those obtained by humans. As a consequence, many have attributed superhuman performance to such systems, and some tasks have been deemed solved. However, while systems in some areas of AI are compared with the best possible humans, e.g. IBM Deep Blue vs. Garry Kasparov in chess² or IBM Watson vs. Ken Jennings and Brad Rutter in the *Jeopardy!* quiz show³, NLP researchers often naively or vaguely estimate the “human baseline”, assuming it is a uniform and accepted term of comparison, an established level that systems need to simply beat. In this section we provide a broad overview of existing NLP benchmarks, with a particular focus on NLU leaderboards where human baselines are outperformed by systems, and then show that the construction of such benchmarks is fraught with inconsistencies.

The SuperGLUE benchmark (Wang et al., 2019) is a well-known framework for evaluating research towards general-purpose language understanding models for English. It is a collection of 10 language understanding tasks built on existing public datasets, together with private test data, an evaluation server, and a single-number performance metric. In many tasks, humans are outperformed by the best-scoring systems, often by a large margin, ranking 8th in the current overall leaderboard. Likewise, the SuperGLUE predecessor, i.e. GLUE (Wang et al., 2018), was built to measure advances in NLU, and the systems’ scores quickly saturated the benchmark, thereby sliding the human baseline down to the 23rd position in the ranking.

The RACE benchmark (Lai et al., 2017) was designed specifically for evaluating NLP models on a set of challenging reading comprehension tasks, such as Question Answering (QA) and text summarization. It consists of a large dataset of more

that Natural Language Generation (NLG) tasks such as machine translation, where the most accepted ‘leaderboards’ are based on human scoring, have their own set of issues (e.g. Läubli et al., 2020).

²https://en.wikipedia.org/wiki/Deep_Blue_versus_Garry_Kasparov

³https://en.wikipedia.org/wiki/IBM_Watson#Jeopardy!

than 28,000 multiple-choice questions, which are drawn from middle and high school problems extracted from English examinations in China. These questions cover a wide range of topics and require the ability to reason, understand context, and make inferences based on the provided text. Human baselines rank 21st on the public leaderboard, with a gap of almost 20 points compared to the best-scoring system. Similarly, the SQuAD2.0 benchmark (Rajpurkar et al., 2018) is another popular collection of reading comprehension questions and answers based on Wikipedia articles. The questions are created by crowdworkers and the answers are a portion of text from the corresponding article. The peculiar difference of this benchmark compared to SQuAD1.1 (Rajpurkar et al., 2016) is that some of the questions may not have answers, hence systems are required to learn to abstain as well. Again, the human baseline is placed in low positions of the ranking, reaching just the 30th place. Another notable, related benchmark is CoQA (Reddy et al., 2019), a large-scale dataset focused on Conversational QA systems. In this task, humans rank 7th, with a gap of 2 points from the top system.

Quite different results are observed when moving to a cross-lingual scenario or when systems are required to perform mathematical and logical reasoning. In particular, XTREME (Hu et al., 2020) is a benchmark for cross-lingual transfer evaluation that covers dozens of languages spanning 12 language families, and includes 9 tasks that require reasoning about different levels of syntax or semantics. In this case, the human baselines beat the systems in all tasks, with an overall score 8 points higher than that of the best-performing system. XTREME has been succeeded by XTREME-R (Ruder et al., 2021), which covers 14 language families and includes 10 tasks. Similarly, when systems are evaluated over MathQA (Amini et al., 2019) inputs, i.e. mathematical questions in the form of natural language, systems perform poorly compared to humans. Indeed, humans achieve an accuracy of 87% while systems only reach 54.2%. Since systems are still far from human-level performance in these benchmarks, they are out of the scope of our study. However, the highlighted gaps should encourage further research in these areas.

An alternative view on system evaluation is presented by the adversarial evaluation framework (Nie et al., 2020; Kiela et al., 2021), where the evaluation is performed through an iterative “human-

and-model-in-the-loop” annotation process. Humans are asked to inspect the model output and produce adversarial examples that target specific model weaknesses. The evaluation target is thus a moving goalpost, as opposed to the static targets of most other benchmarks, which saturate quickly. The Dynabench benchmark (Kiela et al., 2021) embraces this approach, incorporating tasks such as NLI, QA, sentiment analysis and hate speech detection. It provides a platform for the annotators to examine model output and create adversarial examples. At the time of writing, most of the tasks within Dynabench do not report human performance, however. Exceptions include the adversarial visual QA task (Sheng et al., 2021), where the proposed adversarial examples are solved by other humans and agreement is computed in terms of accuracy. Model performance in this setting falls far below the human performance.

Using more challenging examples for model evaluation, and possibly subsequent re-training, is an appealing approach, likely to strengthen the models with respect to the aspects that the examples target. The caveat is, however, that special care needs to be taken to avoid loss of generality. The annotation of adversarial examples directly depends on the behavior of the model (or set of models) under consideration; the addition of a large number of adversarial examples will likely change the data distribution by potentially overemphasizing rare events; finally, the annotators may focus on a small number of properties of the model, thus “overfitting” the models.

Although there are many other popular NLP benchmarks to be investigated, e.g. XGLUE (Liang et al., 2020) and SentiBench (Ribeiro et al., 2016), we limit our review to those in which human performance is provided and that can therefore help us answer the main question of this paper concerning the meaning of superhuman performance.

3 Human Baselines Are Not Reliable

As discussed above, many NLU benchmarks are saturated (cf. Figure 1). Here we dive deeper into some of them, identify the reasons for their quick saturation, and discuss whether it is fair to claim superhuman performance of state-of-the-art models. In particular, we study SuperGLUE (Wang et al., 2019) and SQuAD (Rajpurkar et al., 2016, 2018), as the representatives for general language understanding and reading comprehension, respectively.

3.1 SuperGLUE

For each of the ten tasks in SuperGLUE, human performance is provided and systems are compared against it. Specifically, for four of these tasks – Word in Context (WiC, Pilehvar and Camacho-Collados, 2019), Multi-Sentence Reading Comprehension (MultiRC, Khashabi et al., 2018), Recognizing Textual Entailment (RTE, Nangia and Bowman, 2019), Reading Comprehension with Commonsense Knowledge (ReCoRD, Zhang et al., 2018) – human performance is computed by the authors of the corresponding papers, while for the remaining tasks⁴ humans are evaluated by the creators of the SuperGLUE benchmark.

WiC For this lexical-semantic task, four sets of 100 instances with an overlap of 50 instances between two of the annotators were randomly sampled from the test set. Each set was then assigned to an annotator, resulting in a total of 300 annotated instances. The annotators were not lexicographers and were not provided with sense distinctions to resemble the more difficult scenario for unsupervised models (cf. Appendix C). A final score of 80% was then obtained by averaging the individual scores achieved by the humans on the 4 sets (between 79% and 82%).

MultiRC In the Multi-Sentence Reading Comprehension task, four native-speaker annotators tagged the entire test set of 166 instances. Human performance was obtained by combining the individual predictions of the different annotators via majority voting.

RTE To establish the human performance on the RTE task, annotators were hired through the Hybrid data collection platform. Each annotator first completed a short training procedure, during which they were provided with task-specific guidelines and annotated 20 random examples from the dev set. Only annotators with $\geq 65\%$ accuracy qualified for the main task. 500 examples were randomly taken from the test set and, for each instance, the final label was obtained by combining 5 different annotations via majority voting, reporting a final accuracy of 93.6%. The average pay rate was \$17/hr for the main task, and \$7.6/hr for training.

⁴Boolean Questions (BoolQ), Commitment Bank (CB), Choice of Plausible Alternatives (COPA), Winograd Schema Challenge (WSC), Broadcoverage Diagnostics (AX-b), Winograd Schema Diagnostics (AX-g).

ReCoRD For the Reading Comprehension with Commonsense Knowledge task, 2,257 crowdworkers were hired through the Amazon Mechanical Turk platform (AMT). For first-time workers, the HIT⁵ assignments were accompanied with guidelines. Crowdworkers were required to have ≥ 50 HITs with a 95% HIT acceptance rate and to be located in the USA, Canada, or UK. The average pay rate was \$3.6/hr.

Other SuperGLUE Tasks For the six remaining tasks, the SuperGLUE authors hired crowdworkers through AMT: the annotators first completed a short training phase where 30 random development set examples were provided for each task. Only workers who completed 5 HITs during training with performance at, or above, the median across all workers were admitted to the main task. Human performance was estimated on a random set of 100 test samples from each task, by applying majority voting on the annotations of 5 workers. During both phases, workers had access to task-specific guidelines, with a pay rate of \$23.75/hr.

3.2 SQuAD

In SQuAD1.1 (Rajpurkar et al., 2016), the researchers obtained ≥ 3 answers from human workers for each question in the dev and test sets, and estimated human performance by using only one of the answers as the “human prediction” and the remaining answers as “ground truth” for comparison. Specifically, workers were shown the questions and relevant paragraphs of an article and were asked to select the shortest paragraph span that answered the question. They were advised to complete 5 questions in 2 minutes with a \$9/hr pay rate.

In SQuAD2.0 (Rajpurkar et al., 2018), instead, the authors collected multiple answers for each question (i.e. 4.8 answers, on average) and selected the final human prediction by majority voting. The answers were collected by providing annotators with a paragraph and its associated questions – unanswerable and answerable ones shuffled together – and asking them either to highlight the answer in the paragraph or to mark the question as unanswerable. They were asked to spend one minute per question with a \$10.50/hr pay rate.

⁵A Human Intelligence Task, or HIT, is a question that needs an answer. A HIT represents a single, self-contained, virtual task that a worker can work on, submit an answer, and collect a reward for completing.

Model	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b
VEGA v2	90.5	98.6	99.4	88.2	94.4	96.0	77.4	98.6	100.0	-0.4
ST-MoE-32B	92.4	96.9	99.2	89.6	95.1	93.5	77.7	96.6	96.1	72.3
TURING NLR v5	92.0	95.9	98.2	88.4	96.4	94.1	77.1	97.3	93.3	67.8
ERNIE 3.0	91.0	98.6	97.4	88.6	94.7	92.6	77.4	97.3	92.7	68.6
PALM 540B	91.9	94.4	99.0	88.7	94.2	94.1	77.4	95.9	95.5	72.9
HUMAN BASELINES	89.0	95.8	100.0	81.8	91.7	93.6	80.0	100.0	99.3	76.6

Table 1: Results on SuperGLUE (from <https://super.gluebenchmark.com/leaderboard>). On top, we report the results of the 5 best-performing models, on average, while on bottom we report those of the human baselines on the various tasks. We mark in bold the best scores per task. AX-g and AX-b are the two diagnostic datasets.

3.3 Issues

Comparing the performance of the five best systems against humans on SuperGLUE (Table 1), it is immediately apparent that the machines outperform humans on 6 out of 10 tasks, and often by a large margin (e.g. 7.8 F_1 points on MultiRC). Similarly, best systems substantially outperform humans on SQuAD1.1 and SQuAD2.0, with a margin of 8.3 and 4.1 points in exact match accuracy, respectively. Interestingly, (zero-shot) ChatGPT performs poorly compared to both human baselines and best-performing (fine-tuned) systems. Indeed, compared to the scores reported in Table 1, it achieves just 86.8 on BoolQ, 89.3 on CB, 58.0 on COPA, 85.2 on RTE and 64.6 on WiC as measured by Qin et al. (2023) and Kocoń et al. (2023). Additionally, Kocoń et al. (2023) also showed that ChatGPT performs 20% worse than state-of-the-art systems on the SQuAD2.0 benchmark, and demonstrated that it is, on average, 25% worse than specialized ML systems on a wide array of tasks. Hence it is not relevant for our study as its performance is still far from human-level.

What does appear relevant, instead, are the extremely high, often superhuman, scores achieved by specialized systems. Nevertheless, notwithstanding such scores, in the above-mentioned benchmarks multiple factors make human-to-system comparisons unfair because they limit human performance while facilitating systems. We list them in the remainder of this section.

Apples and oranges The most glaring problem is that, on almost all SuperGLUE tasks, humans and systems are evaluated on different test sets (i.e. on a small subset vs. the full test set). Specifically, in the WiC and RTE tasks, humans are assessed on 21.4% and 16.6% of the test set (i.e. 300 out of 1400 and 500 out of 3000 instances), respectively. Similarly, in the other SuperGLUE tasks humans are evaluated on a subset of 100 instances per task, which – in the worst case of the BoolQ dataset –

amounts to just 3% of the test set. We provide more details in Appendix B.

Human evaluation metrics Different metrics are used to assess humans across tasks. While most of the tasks employ majority voting, WiC merely averages the scores achieved by humans on 4 small distinct subsets. In SQuAD1.1, humans are evaluated by comparing the tags of an arbitrary annotator against those of two other “ground truth” annotators, thereby likely underestimating the final score.

Heterogeneous and unknown pay rates Pay rates varied considerably across the various tasks, ranging from undeclared pay rates to \$23.75/hr. Low and mediocre wages, as in ReCoRD and SQuAD, may have contributed to suboptimal human performance: the \$3.6/hr pay rate on ReCoRD could be one of the reasons for the large gap between systems and humans, while the unknown pay rate for MultiRC might explain the 18.2% human error rate on this binary classification task.

Ground-truth data quality We identified several errors and ambiguous instances in the gold-standard datasets, some of which we report in Table 2. Importantly, we note that, while systems can find spurious correlations between training and evaluation instances, and therefore provide the correct answer without clear evidence, humans cannot find such correlations, or otherwise may genuinely disagree on what the correct answer is. We elaborate on this point in Appendix A, by analyzing several examples per task, as well as in Appendix C, where we report the results of an ad hoc study concerning the WiC dataset.

Information about annotators and instructions Details of the annotator pool (e.g. the number of annotators, their background and nationality, etc.) are often omitted. Similarly, the absence of training instructions and task guidelines raises questions about the quality of the training phase, if any.

BoolQ	<p>Passage: <i>Shower gel – Shower gels for men may contain the ingredient menthol, which gives a cooling and stimulating sensation on the skin, and some men’s shower gels are also designed specifically for use on hair and body. Shower gels contain milder surfactant bases than shampoos, and some also contain gentle conditioning agents in the formula. This means that shower gels can also double as an effective and perfectly acceptable substitute to shampoo, even if they are not labelled as a hair and body wash. Washing hair with shower gel should give approximately the same result as using a moisturising shampoo.</i></p> <p>Question: <i>is it bad to wash your hair with shower gel</i> Answer: TRUE</p>
CB	<p>Premise: <i>A: I do too, so she couldn’t possibly turn them out like some of these popular writers, B: Huh-uh. A: but oh, her books are just incredible. I don’t think they’ve ever made a movie, do you?</i></p> <p>Hypothesis: <i>they’ve ever made a movie</i> Entailment: FALSE</p>
MultiRC	<p>Paragraph: <i>What causes a change in motion? The application of a force. Any time an object changes motion, a force has been applied. [. . .] It depends on the strength of the force. It also depends on the objects mass. Think about some simple tasks you may regularly do. You may pick up a baseball. This requires only a very small force.</i></p> <p>Question: <i>What factors cause changes in motion of a moving object?</i> Candidate Answers: <i>Shape of the object (FALSE), Mass of the object (TRUE), The object’s mass (FALSE), . . .</i></p>
RTE	<p>Premise: <i>In most Pacific countries there are very few women in parliament.</i></p> <p>Hypothesis: <i>Women are poorly represented in parliament.</i> Entailment: TRUE</p>
WiC	<p>Context 1: <i>The senator received severe criticism from his opponent.</i></p> <p>Context 2: <i>The politician received a lot of public criticism for his controversial stance on the issue.</i> Sense Match: FALSE</p>

Table 2: Some errors and ambiguous instances we have found in the gold standard datasets of various SuperGLUE tasks. We limited our analysis to tasks where suspiciously low human performance was reported (cf. Table 1).

4 Setups Favor Misleading Comparisons

Summarizing the above observations, we find four main sources of human-to-system comparison error. These correspond to the following key aspects of the evaluation process: system performance, the evaluation data, the measurement process, and humans themselves. We discuss each in turn.

4.1 Systems: Right for the Wrong Reasons

Across a variety of tasks, Sogaard et al. (2021) report that random train-test splits consistently overestimate model performance: randomization at the sentence level reduces discrepancies between training and test sets as sentences from the same documents occur in both. Non-random standard splits also bring the danger of inadvertent, community-wide overfitting (Gorman and Bedrick, 2019)⁶.

In natural language inference (NLI), multiple authors have found that BERT achieves what looks like near-human accuracy by exploiting idiosyncrasies of the data: they are “right for the wrong reasons” (McCoy et al., 2019; Niven and Kao, 2019). Here much of BERT’s success is attributed to its ability to learn syntactic and lexical cues for inference, which happen to be mostly correct on the original test data. However, these cues do not actually support such inferences on adversarial datasets, taking BERT’s accuracy to chance level or below.

Poliak et al. (2018) report an even more extreme

case of being “right for the wrong reason”: several NLI datasets support what they call hypothesis-only models, which perform surprisingly well without exposure to the premise (Gururangan et al., 2018), e.g. outperforming the majority-class baseline. Poliak et al. (2018) attribute this to statistical irregularities in the data (often single words indicating negation), caused by obvious annotation strategies chosen by crowdworkers who were not stimulated enough to come up with more creative ways to produce contradictions or entailments. Along the same lines, Parmar et al. (2023) recently identified instruction bias in 14 NLU benchmarks. Specifically, they found that this phenomenon is evident in most of these datasets, showing that $\sim 73\%$ of instruction examples, on average, share a few clear bias patterns, and that models often fail to generalize beyond such patterns.

4.2 Data: Monolithicity Obscures Details

A further cause of systematic performance overestimation is that test sets include instances with varied, often unfathomable, levels of difficulty, so the exact reported accuracy will be a weighted average that depends directly on the mixture of easy and hard instances in the test data. The composition of train-test splits can thus make a big difference (Swayamdipta et al., 2020).

In QA, Lewis et al. (2021) investigated the train-test splits of several popular datasets. They found that there can be substantial overlap between the answers and even the questions of the training and test sets. The evaluation results differed greatly be-

⁶Even with hidden test sets, such overfitting could happen via publication bias or just the faster spread of methods used in “state-of-the-art” systems (as measured on a static test set).

tween seen and unseen questions and answers; for instance, the exact-match accuracy of BART as a closed-book QA system on WebQuestions dropped from 76% to below 2% when neither the question nor the answer were ever seen during training.

In semantic parsing, seq2seq models such as BART and T5 are very accurate when evaluated in-domain on broad-coverage parsing tasks, e.g. Bevilacqua et al. (2021a). Yao and Koller (2022) report that their accuracy drops to close to zero on test subsets that require them to generalize to language that is structurally more complex than the training data. This is corroborated when constructing hard sets, i.e. train-test splits based on compositional generalization, forcing the accuracy of seq2seq models below 20% (Bogin et al., 2022).

4.3 Measurement: Automation Is Limiting

A third key limitation of current evaluations, and especially existing leaderboards, is that they assume that the performance of a model can be measured automatically. While this has not been discussed very much in NLU, in other communities it has long been recognized that automatic evaluations are imperfect proxies of human judgments (Novikova et al., 2017). Machine translation papers report BLEU scores because they are drastically cheaper to calculate than the cost to collect human judgments about the fluency and adequacy of text; but one system that outperforms another on BLEU is not necessarily judged better by humans (Callison-Burch et al., 2006; Popel et al., 2020). While recent automatic metrics correlate better with human judgments (Kocmi et al., 2021), automatic evaluation has consistently been found problematic when comparing top-performing systems (Ma et al., 2019). Similarly, Byron et al. (2009) recommend crowd-sourced evaluations to counter the inadequacy of automated evaluation for NLG.

The deeper issue with our reliance on automated evaluations is that they constrain the tasks on which we can evaluate systems. New shared tasks and datasets are specifically designed to make automated evaluations possible. However, many skills that show competent language use cannot easily be approximated by automatic measures (Dunietz et al., 2020): there are entire facets of language competence that are systematically out of scope for the tasks we design. One might argue that these are the most interesting parts of the actual mastery of language. Therefore, human-level performance

on automatically-evaluated tasks does not equate to human-level performance on real language use.

4.4 Humans: They Often Disagree

The final and possibly most problematic issue with system evaluation lies in the creation of the evaluation data itself. Common evaluation methodology assumes that there exists a single ground-truth for evaluation. This is a great oversimplification. We argue that evaluation should be conducted with reference to different groups of annotators to go beyond a one-dimensional performance score, to reflect multiple possible ‘truths’.

A great deal depends on how annotators are instructed to produce the data. It is well-known that human annotation quality may suffer from errors resulting from lack of attention given to the task, both by annotators themselves and by the annotation managers, often resulting from the need to drive annotation costs down (Mishra and Gorana, 2021). Importantly, however, human label variation does not always reflect poor annotation. Label variation can also result from stimulus characteristics or the context in which annotation occurs, including factors like the identity of the annotators, their background, and world knowledge. Plank (2022) identifies three main reasons for human label variation, namely annotator disagreement, subjectivity (multiple possible perspectives) and underspecification (multiple plausible answers). While subjectivity (e.g., due to cultural differences) is a clear issue in tasks like hate speech detection (Davani et al., 2021), inherent disagreements, ambiguous sentence meaning, underspecification in guidelines and annotator behavior have been identified not only in fine-grained Word Sense Disambiguation tasks (Navigli, 2009), but even in NLI (Pavlick and Kwiatkowski, 2019; Zhang and de Marneffe, 2021; Jiang and de Marneffe, 2022).

While the standard approach for training and evaluating NLP systems is to use a single gold label for each example, a growing body of work deals with multiple labels by varying model training in various ways: different aggregation methods (Paun et al., 2018), training on the distributional labels (Potts et al., 2020), learning from agreement signals (Plank et al., 2014), or modeling the annotators (Geva et al., 2019; Sap et al., 2022; Gordon et al., 2022). Recently, Basile et al. (2021) proposed extending this approach to evaluation. Fully benefiting from this extension requires releasing

annotator characteristics labels (Prabhakaran et al., 2021), including socio-demographic information, and carefully documenting the annotation process (Geburu et al., 2018; Bender and Friedman, 2018; Geiger et al., 2020).

Annotator disagreement often results from differences across individuals – not just in NLP but also in fields such as cognitive science (Levinson, 2012) and psycholinguistics (Kidd et al., 2018). This phenomenon is often underestimated, since experiments tend to focus on a homogeneous sub-sample of the human population (Henrich et al., 2010). Annotators have different natural biases (Reidsma and op den Akker, 2008), and models often learn annotator-specific signals that are not generalizable (Geva et al., 2019), including opinion, personality (Sap et al., 2022) and culture (Hershcovich et al., 2022), but also different interpretation of guidelines (Hansen and Søgaard, 2021; Parmar et al., 2022). To deal with subjectivity, Rottger et al. (2022) recently introduced two contrasting data annotation paradigms: the descriptive and prescriptive ones. While the former encourages annotator subjectivity by capturing and modelling different beliefs, the latter, instead, discourages it and enforces annotators to encode one specific belief, formulated in the annotation guidelines. Depending on the downstream application of the dataset, one paradigm can be more suitable than the other, but neither paradigm is inherently superior. However, dataset annotators should explicitly aim for one of the two paradigms to facilitate the intended downstream use of their dataset, and to document, for the benefit of others, how exactly their dataset was annotated.

In conclusion, without more attention to the “science of annotation”, the methodological laxity in today’s dataset creation will continue to foster inaccurate estimations of human performance.

5 Humans Can Explain Their Answers

When performing language tasks, humans are capable of explaining why they provided a given answer. Thus, when models are claimed to attain human-level language understanding, we can reasonably expect to be able to elicit explanations from them. This has proven highly challenging, however, which casts further doubts on such claims.

Why do we need explanations? At the level of an individual problem instance, explanations can help users assess whether to trust a given answer. At the level of a system, they help regulators and

the general public to assess whether, or in what contexts, a system is safe to use, e.g. by uncovering unwanted biases or by revealing that the system relies on outdated knowledge. In the context of this paper, explanations can help NLP researchers understand the behaviour of their systems, e.g. to make sure that models are right for the right reasons (McCoy et al., 2019; Niven and Kao, 2019), or to uncover some of the shortcuts that the model may have learned (Geirhos et al., 2020), as discussed in §4.1. Indeed, the absence of explanations can lead researchers astray. For example, in a prize-winning paper, Kaushik and Lipton (2018) analysed several state-of-the-art QA systems and found that they simply classified the best matching answer using their pre-stored knowledge about each question candidate, without performing any ‘reading’. None of the papers in which these QA systems were introduced had considered this possibility.

What are the challenges? While the importance of explanations is well-understood, progress has been hampered by various issues. One issue is that the evaluation of system-generated explanations is hard to automate (§4.3). Another issue is that it is not always clear what form the explanations should take. For tasks such as sentiment classification, it may be sufficient to highlight which words from the input text have mostly affected a given prediction. However, for NLI and QA, providing informative explanations can be challenging, even for humans. This can be observed by inspecting datasets that include human explanations (Camburu et al., 2018; Rajani et al., 2019; Aggarwal et al., 2021). Finally, system-generated explanations are typically not faithful, i.e. they do not necessarily reflect the process used by the model. For instance, Camburu et al. (2020) found that models can generate contradictory explanations for a given input.

6 Recommendations

Based on the findings of the previous sections, we argue that current claims regarding superhuman performance are not adequately grounded, leading to unjustified hype. Here we provide a set of recommendations aimed at making comparisons between humans and machines fairer and more reliable.

Do not favor machines against humans Various actions can be taken to set a level playing field between humans and machines, so as to provide a more realistic sense of their actual performance:

1. **Avoid using the same documents for training and evaluation** (§4.1): in fact, using the same documents inherently reduces discrepancies across splits (Gorman and Bedrick, 2019), encouraging models to learn specific idiosyncrasies that appear in both (McCoy et al., 2019).
2. **Balance easy and hard test set items** (§4.2), so as to report accuracies and enable analyses based on their difficulty level.
3. **Occasionally refresh test sets** (§2), as suggested by recent trends in adversarial evaluation (Kiela et al., 2021).
4. **Complement automatic evaluations with human judgements** (§4.3), so as to compare systems with humans on facets of language use that cannot be evaluated automatically.
5. **Adequately train and motivate humans** (§3.3), aiming to increase the quality of human annotations through a solid training process and higher pay, in a sense mimicking the effort taken in improving systems.

Make human performance evaluation transparent and reproducible We suggest carrying out an effort similar to systems’ reproducibility for evaluating humans as well, including:

1. **Document the annotator pool composition** (§3.3), by explicitly answering the following questions: how many annotators were hired? Following what process? What is their cultural background, nationality, languages and areas of expertise? What is their hourly pay rate?
2. **Specify the annotation process** (§3.3 and §4.4): it is important to state how many annotators were assigned to each instance, the training process they underwent, the guidelines they received (and how such guidelines were fine-tuned), and the metrics used to compute the overall human performance (averaging individual scores, majority voting, etc.).
3. **Provide individual annotations** (§4.4): this allows recalculation of overall human performance whenever new metrics are tried, identifying the best metrics, calculating the scores of the best and worst annotators, the gap between the two, and the correlation between

metrics and individual annotators – all aspects that the annotation community has long advocated. Importantly, the availability of individual answers, combined with the annotators’ profiles, opens the door to deeper investigations about why and when humans disagree.

Increase annotation accountability Multiple measures can be implemented to make both systems and benchmarks more reliable, transparent and informative:

1. **Include explanations in your benchmark** (§5): requiring annotators to provide the rationale behind their choices implicitly enforces them to devote more attention to the annotation task, thus yielding higher-quality and more consistent annotations. Moreover, annotators’ explanations can be used to study subjectivity, and discover (and mark) ambiguous instances.
2. **Let systems produce explanations** (§5): before claiming superhuman performance, it is important that, similarly to humans, systems can explain the inferences behind their predictions. This is key both for increasing systems’ credibility and for discovering their limitations. However, it is not impossible that a system will produce the right answer with the wrong explanation, or vice versa. For this reason, we believe that a system must be able to provide explanations that support its answers without knowing that answer a priori, inferring the answer based on its knowledge.

7 Conclusions

We have discussed the distressing tendency of many NLP researchers to claim superhuman performance for their systems, and outlined why such claims are not (yet) grounded. We identified problems with evaluation data, evaluation measures and methodology, system understandability, and the human creation of data, all of which contribute to our conclusion.

As a final remark, with this paper we hope to make the reader more suspicious and rigorous when claims about “superhuman” performance are made, and, more importantly, to incentivize benchmark creators to address current limitations and design more solid and transparent benchmarks that will advance our scientific understanding of NLP systems and humans.

8 Limitations

In this paper, we have unearthed a variety of problems present in current evaluation benchmarks that favor systems over humans, or that simply make such comparisons unfair. We conclude that there is no real evidence to claim that today’s language models possess superhuman performance. However, without empirical results obtained under the right setups, we cannot even claim the opposite, namely that humans are still better than systems. We leave such demonstrations for future work.

Additionally, while a good portion of the NLP research effort is devoted to natural language generation (NLG) tasks (which includes MT), here we provide only some pointers to NLG/MT. Indeed, as discussed in Section 4.3, these problems exist in the NLG universe as well, but, due to space constraints, we limit our analysis to NLU tasks.

Acknowledgments

This work grew out of a brainstorming session held at the Rome Workshop on Ten Years of BabelNet in July 2022.⁷

We gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the  European Union’s Horizon 2020 research and innovation programme, and the support of the PNRR MUR project PE0000013-FAIR. 

This work has been carried out while Simone Tedeschi was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome. The contribution of Jan Hajič has been supported by the LUSyD project No. GX20-16819X, funded by the Czech Science Foundation, and has used resources provided by the LRI LINDAT/CLARIAH-CZ, project LM2023062 funded by the MŠMT CR. The DFKI contribution to this work is supported by the LT-BRIDGE project, which has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 952194. Rico Sennrich was funded by the Swiss National Science Foundation (grant no. 176727).

⁷<http://mousse-project.org/events/event-a5f3r5.html>

References

- Shourya Aggarwal, Divyanshu Mandowara, Vishwa-jeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. [Explanations for CommonsenseQA: New Dataset and Models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. [MathQA: Towards interpretable math word problem solving with operation-based formalisms](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.
- BT Sue Atkins and Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. [We need to consider disagreement in evaluation](#). In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021a. [One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline](#). In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-21)*, volume 35, pages 12564–12573. AAAI Press.
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021b. [Recent trends in word sense disambiguation: A survey](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Ben Bogin, Shivanshu Gupta, and Jonathan Berant. 2022. [Unobserved local structures make compositional generalization hard](#). In *Proceedings of EMNLP*.
- Donna Byron, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2009. [Report on the First NLG](#)

- Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 165–173, Athens, Greece. Association for Computational Linguistics.
- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.
- Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations.
- Jesse Duniety, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal III Daumé, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.
- R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- 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.
- Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324.
- Victor Petré Bach Hansen and Anders Søgaard. 2021. Guideline bias in Wizard-of-Oz dialogues. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 8–14, Online. Association for Computational Linguistics.
- Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarelli, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *arXiv preprint arXiv:2209.03392*.
- Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In

- Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.
- Daniel Khoshdel, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. [Looking beyond the surface: A challenge set for reading comprehension over multiple sentences](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
- Evan Kidd, Seamus Donnelly, and Morten H Christiansen. 2018. [Individual differences in language acquisition and processing](#). *Trends in cognitive sciences*, 22(2):154–169.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ring-shia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2022. [Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future](#). *Computational Linguistics*, pages 1–42.
- Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. [To ship or not to ship: An extensive evaluation of automatic metrics for machine translation](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.
- Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniec, Marcin Gruga, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. [Chatgpt: Jack of all trades, master of none](#).
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. [A Set of Recommendations for Assessing Human–Machine Parity in Language Translation](#). *Journal of Artificial Intelligence Research (JAIR)*, 67:653–672.
- Stephen C Levinson. 2012. [The original sin of cognitive science](#). *Topics in cognitive science*, 4(3):396–403.
- Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. [Question and answer test-train overlap in open-domain question answering datasets](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fen-fei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. [XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.
- Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. [Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.
- Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. [SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation \(MCL-WiC\)](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Abhilash Mishra and Yash Gorana. 2021. [Who decides if AI is fair? the labels problem in algorithmic auditing](#).
- Nikita Nangia and Samuel R. Bowman. 2019. [Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

- Roberto Navigli. 2009. [Word sense disambiguation: A survey](#). *ACM Comput. Surv.*, 41(2):10:1–10:69.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Timothy Niven and Hung-Yu Kao. 2019. [Probing neural network comprehension of natural language arguments](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. [Why we need new evaluation metrics for NLG](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.
- Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. [Don’t blame the annotator: Bias already starts in the annotation instructions](#). *arXiv preprint arXiv:2205.00415*.
- Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. [Don’t blame the annotator: Bias already starts in the annotation instructions](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1779–1789, Dubrovnik, Croatia. Association for Computational Linguistics.
- Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. [Comparing Bayesian models of annotation](#). *Transactions of the Association for Computational Linguistics*, 6:571–585.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. [WiC: the word-in-context dataset for evaluating context-sensitive meaning representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- Barbara Plank. 2022. [The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation](#). *arXiv preprint arXiv:2211.02570*.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. [Learning part-of-speech taggers with inter-annotator agreement loss](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. [Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals](#). *Nature Communications*, 11(1):4381.
- Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. [DynaSent: A dynamic benchmark for sentiment analysis](#).
- 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.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. [Is chatgpt a general-purpose natural language processing task solver?](#)
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Explain yourself! leveraging language models for commonsense reasoning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [CoQA: A conversational question answering challenge](#). *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Dennis Reidsma and Riëks op den Akker. 2008. [Exploiting ‘subjective’ annotations](#). In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.
- Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. [Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods](#). *EPJ Data Science*, 5(1):1–29.
- Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. [Evaluation examples are not equally informative: How should that change NLP leaderboards?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. [XTREME-R: Towards more challenging and nuanced multilingual evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. [Human-adversarial visual question answering](#).
- Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. [We need to talk about random splits](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. [Dataset cartography: Mapping and diagnosing datasets with training dynamics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Yuekun Yao and Alexander Koller. 2022. [Structural generalization is hard for sequence-to-sequence models](#). In *Proceedings of EMNLP*.
- Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. [Record: Bridging the gap between human and machine commonsense reading comprehension](#).
- Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. [Identifying inherent disagreement in natural language inference](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

A Ground Truth Data Quality

In Table 2, we reported one problematic example for each of the most crucial tasks in the SuperGLUE benchmark. Here we comment on those examples and provide additional problematic cases which we identified by manually inspecting the datasets. Some of these cases appear recurrently.

BoolQ The example in Table 2 is blatantly wrong, as it explicitly says that shower gel is an *effective and perfectly acceptable substitute to shampoo*, hence the label should be FALSE. We provide more errors in Table 3. Specifically, we believe

Passage: *Columbidae* – The distinction between “doves” and “pigeons” is not consistent. In modern everyday speech, as opposed to scientific usage or formal usage, “dove” frequently indicates a pigeon that is white or nearly white. However, some people use the terms “dove” and “pigeon” interchangeably. In contrast, in scientific and ornithological practice, “dove” tends to be used for smaller species and “pigeon” for larger ones, but this is in no way consistently applied. Historically, the common names for these birds involve a great deal of variation between the terms. The species most commonly referred to as “pigeon” is the species known by scientists as the rock dove, one subspecies of which, the domestic pigeon, is common in many cities as the feral pigeon.

Question: *is there a difference between doves and pigeons* **Answer:** FALSE

Passage: *Redback spider* – The redback is one of the few spider species that can be seriously harmful to humans, and its liking for habitats in built structures has led it to being responsible for a large number of serious spider bites in Australia. Predominantly neurotoxic to vertebrates, the venom gives rise to the syndrome of latrodectism in humans; this starts with pain around the bite site, which typically becomes severe and progresses up the bitten limb and persists for over 24 hours. Sweating in localised patches of skin occasionally occurs and is highly indicative of latrodectism. Generalised symptoms of nausea, vomiting, headache, and agitation may also occur and indicate severe envenomation. An antivenom has been available since 1956. There have been no deaths directly due to redback bites since its introduction, however Isbister et al. have suggested patients for whom antivenom is considered should be fully informed “there is considerable weight of evidence to suggest it is no better than placebo”, and in light of a risk of anaphylaxis and serum sickness, “routine use of the antivenom is therefore not recommended”. As of the 2013 (updated 2014) edition of the *Snakebite & Spiderbite Clinical Management Guidelines from NSW HEALTH* (latest available in 2017), Red-back spider bites were considered not life-threatening but capable of causing severe pain and systemic symptoms that could continue for hours to days.

Question: *can a red back spider bite kill you* **Answer:** TRUE

Passage: *Mobile phones in prison* – In most prisons, inmates are forbidden from possessing mobile phones due to their ability to communicate with the outside world and other security issues. Mobile phones are one of the most smuggled items into prisons. They provide inmates the ability to make and receive unauthorized phone calls, send email and text messages, use social media, and follow news pertaining to their case, among other forbidden uses.

Question: *are you allowed to have your phone in prison* **Answer:** FALSE

Passage: *Vena amoris* – *Vena amoris* is a Latin name meaning, literally, “vein of love”. Traditional belief established that this vein ran directly from the fourth finger of the left hand to the heart. This theory has been cited in western cultures as one of the reasons the engagement ring and/or wedding ring was placed on the fourth finger, or “ring finger”. This traditional belief is factually inaccurate as all the fingers in the hand have a similar vein structure.

Question: *is it true that the ring finger is connected to the heart* **Answer:** FALSE

Passage: *Substitute (association football)* – Most competitions only allow each team to make a maximum of three substitutions during a game and a fourth substitute during extra time, although more substitutions are often permitted in non-competitive fixtures such as friendlies. A fourth substitution in extra time was first implemented in recent tournaments, including the 2016 Summer Olympic Games, the 2017 FIFA Confederations Cup and the 2017 CONCACAF Gold Cup final. A fourth substitute in extra time has been approved for use in the elimination rounds at the 2018 FIFA World Cup, the UEFA Champions League and the UEFA Europa League. Each team nominates a number of players (typically between five and seven, depending on the competition) who may be used as substitutes; these players typically sit in the technical area with the coaches, and are said to be “on the bench”. When the substitute enters the field of play it is said they have come on or have been brought on, while the player they are substituting is coming off or being brought off.

Question: *can a player be substituted twice in football* **Answer:** TRUE

Table 3: Additional problematic instances we have found in the BoolQ dataset.

that some of these examples are wrongly annotated, ambiguous, or highly misleading. In the first example, from the premise, it seems that *some scientists and ornithologists differentiate between doves and pigeons*, so the answer might be subjective, and therefore ambiguous. In the second example, instead, it seems there is no evidence that a red back spider bite can kill a human being, but the answer is TRUE. Similarly to the first case, in the third example the premise states that *in most prisons* possession of mobile phones is not allowed, thus the answer might change depending on the prison. In the fourth example, the fact that *all the fingers have a similar vein structure* does not mean that *the ring finger is not connected to the heart*, on the contrary, this reinforces the hypothesis. Finally, while

two or more players can be substituted in a football game, the same player cannot be substituted twice.

CommitmentBank In the CB example reported in Table 2 we have that A does not know whether *they’ve ever made a movie* and, indeed, asks if B thinks they have. Therefore, we cannot conclude that *the movie was never made*, and the answer should be NEUTRAL. By inspecting the dataset, we discovered that its instances follow standard patterns that can be easily learned by machines, but, at the same time, confuse humans. Indeed, most of the time, the entailment is TRUE when a fragment of the hypothesis appears (as an exact match) in the premise (see the second and third examples in Table 5). To the contrary, the entailment is FALSE when

Paragraph: Amateur tennis star Guy Haines wants to divorce his vulgar and unfaithful wife Miriam, so he can marry the elegant and beautiful Anne Morton, daughter of a senator. While on a train to meet Miriam, Haines meets Bruno Anthony, a forward stranger who recognizes Guy from gossip items in the newspapers that detail his marital problems. During lunch in Bruno's compartment, Bruno tells Guy about his idea for the perfect "Criss-cross" murder: he will kill Miriam and in exchange, Guy will kill Bruno's father. Since both are strangers, otherwise unconnected, there is no identifiable motive for the crimes, Bruno contends, hence no suspicion. Guy hurriedly leaves the compartment but leaves Bruno thinking he has agreed to the deal. Guy accidentally leaves his cigarette lighter behind, a gift from Anne to Guy, which Bruno pockets. Bruno heads to Guy's hometown of Metcalf and follows Miriam and her two beaux to an amusement park, where he briefly illuminates her face with Guy's lighter, then strangles her to death. Guy's problems begin when his alibi an inebriated college professor on the same train as Guy can not remember their meeting. But they increase exponentially when Bruno makes repeated appearances into Guy's life as he seeks to remind Guy that he is now obliged to kill Bruno's father, according to the bargain he thinks they struck on the train. Bruno sends Guy the keys to his house, a map to his father's room, and a pistol. Soon after, Bruno appears at a party at Senator Morton's house and hobnobs with the guests, much to Guy's apprehension and Anne's increasing suspicion.

Question: Who are the two that Guy and Bruno are planning to murder? **Candidate Answers:** Bruno's father (TRUE), Guy's father (FALSE), Bruno's wife (FALSE), Miriam and Bruno's father (TRUE), Guy's wife (TRUE), . . .

Paragraph: Albert Bandura OC (/baen'dU@r@/; born December 4, 1925) is a psychologist who is the David Starr Jordan Professor Emeritus of Social Science in Psychology at Stanford University. For almost six decades, he has been responsible for contributions to the field of education and to many fields of psychology, including social cognitive theory, therapy and personality psychology, and was also influential in the transition between behaviorism and cognitive psychology. He is known as the originator of social learning theory and the theoretical construct of self-efficacy, and is also responsible for the influential 1961 Bobo doll experiment. Social learning theory is how people learn through observing others. An example of social learning theory would be the students imitating the teacher. Self-efficacy is "The belief in one's capabilities to organize and execute the courses of action required to manage prospective situations." To paraphrase, self-efficacy is believing in yourself to take action. The Bobo Doll Experiment was how Albert Bandura studied aggression and non-aggression in children. A 2002 survey ranked Bandura as the fourth most-frequently cited psychologist of all time, behind B. F. Skinner, Sigmund Freud, and Jean Piaget, and as the most cited living one. Bandura is widely described as the greatest living psychologist, and as one of the most influential psychologists of all time. In 1974 Bandura was elected to be the Eighty-Second President of the American Psychological Association (APA). He was one of the youngest president-elects in the history of the APA at the age of 48. Bandura served as a member of the APA Board of Scientific Affairs from 1968 to 1970 and is well known as a member of the editorial board of nine psychology journals including the Journal of Personality and Social Psychology from 1963 to 1972. At the age of 82, Bandura was awarded the Grawemeyer Award for psychology.

Question: In what year was Bandura awarded the Grawemeyer Award for psychology. **Candidate Answers:** 2010 (FALSE), 2007 (TRUE), 2000 (FALSE), 2002 (TRUE)

Paragraph: (CNN) – German art collector Cornelius Gurlitt, whose nearly priceless collection was confiscated because it was suspected to contain pieces looted by the Nazis, died Tuesday and left the masterpieces to a Swiss museum. One day after Gurlitt's death at the age of 81, the Museum of Fine Arts Bern announced that Gurlitt had named it his unrestricted and unfettered sole heir. The news came as a surprise, the museum said Wednesday, because Gurlitt had never had any connection to it. The museum's directors are delighted at the news, they said in a statement, but also recognize that there are outstanding legal and ethical questions surrounding the collection. Gurlitt had undergone major heart surgery and was hospitalized for many weeks, his representative said in a statement. Gurlitt grabbed the attention of the art world when German prosecutors seized more than 1,200 paintings from his Munich apartment in 2012, including works by Picasso and Matisse. The collection was confiscated as part of an investigation into tax fraud, but then it was thought that some of the paintings may have been works that were looted by the Nazis. Just last month, part of the collection was returned to Gurlitt as part of a deal with Germany's cultural authorities and the Bavarian Justice Ministry. Under the agreement, works owned by Gurlitt that were not under suspicion were returned to him. Those suspected of being stolen were to be held securely while a task force investigates their provenance – and will be returned to their original Jewish owners or their descendants if a claim is proven. Gurlitt's representative said that with the art collector's death, the investigation into the collection ceases. The court that was handling the investigation proceedings will now function as an estate court in the case.

Question: How old was the art collector Cornelius Gurlitt when he died? **Candidate Answers:** At the age of 81 (TRUE), 80 (FALSE), 80 years old (TRUE), 81 (FALSE)

Table 4: Additional problematic instances in the Multi-Sentence Reading Comprehension (MultiRC) dataset.

the same text fragment appears negated either in the premise or in the hypothesis, e.g. preceded by *don't think* or similar, standard constructs (see the first, fourth and fifth examples in Table 5). However, as argued before, the mere fact of not thinking that a thing is true does not necessarily imply that thing is not true.

MultiRC Regarding the MultiRC example of Table 2, in this case, the error is in the candidate

answers. Specifically, two candidate answers are equivalent, i.e. *Mass of the object* and *The object's mass*, but they are labeled differently, namely with TRUE and FALSE tags, respectively. In Table 4 we provide additional errors for this task. Specifically, in the first example, the question explicitly asks "Who are the two that Guy and Bruno are planning to murder?", but the possible answers are i) *Miriam and Bruno's father*, ii) *Bruno's father*

<p>Premise: A: <i>Your turn.</i> B: <i>Okay. Uh, I don't think they should abolish it.</i></p> <p>Hypothesis: <i>they should abolish it</i></p> <p>Entailment: FALSE</p>
<p>Premise: <i>The lunch trade had mostly disappeared so he wasn't hard to spot. He was at a window table but he was ignoring the river, being deep in conversation with a middle-aged man wearing a suit and a short sheepskin car coat with matching brown suede shoes. Even from this distance you could guess the guy's tailor was based in Dublin.</i></p> <p>Hypothesis: <i>the guy's tailor was based in Dublin</i></p> <p>Entailment: TRUE</p>
<p>Premise: B: <i>and, you know, they just love kittens.</i> A: <i>Yeah.</i> B: <i>They just are fascinated.</i> A: <i>Oh, yeah.</i> B: <i>So she doesn't know that this is a cat yet.</i></p> <p>Hypothesis: <i>this is a cat</i></p> <p>Entailment: TRUE</p>
<p>Premise: A: <i>Well, actually, uh, A: I don't think I'm in the, uh, majority in Texas</i></p> <p>Hypothesis: <i>she is in the majority in Texas</i></p> <p>Entailment: FALSE</p>
<p>Premise: B: <i>Because too often, there can be extremism that hurts from any direction, regardless of whatever you're arguing or concerned about.</i> A: <i>Yeah. Right. Yeah, I know, you're right, they would lobby that and I see that, and that's why, you know, I'm like, okay, what's my role in this thing,, you know, what's my part,</i> B: <i>Yeah. A: because I don't think the system is going to get fixed.</i></p> <p>Hypothesis: <i>the system is going to get fixed.</i></p> <p>Entailment: FALSE</p>

Table 5: Additional problematic instances we have found in the CommitmentBank (CB) dataset.

and iii) *Guy's wife*. Although, by design, MultiRC creators explicitly state that multiple answers can be correct, answers are judged independently, so it would not be valid to form a correct answer by combining ii) and iii). These cases are very frequent in MultiRC and might have negatively affected human performance. Furthermore, typos in the questions and/or paragraphs (i.e. *Guty*, in this case) might have further limited their scores.

In the second example, the ground truth answers are “2002” and “2007”. However, while “2007” can be inferred by adding 82 years (i.e. the age at which Albert Bandura received the Grawemeyer award) to his birth date (i.e. 1925), “2002” is a wrong answer. Indeed, the paragraph says that “A 2002 survey ranked Bandura as the fourth most-frequently cited psychologist of all time”, but there is no evidence that he received the award in 2002.

Finally, in the third example, from the paragraph, it is clear that the German art collector Cornelius Gurlitt passed away at the age of 81. However, there are three errors in the possible answers for this entry. First, “*At the age of 81*” and “*81*” are labeled as TRUE and FALSE, respectively. Second, “*80 years old*” is labeled as TRUE, hence contradicting

<p>Premise: <i>Compuware claims that Allan Tortorice and Jim Hildner were among several former employees who revealed trade secrets after they moved to IBM.</i></p> <p>Hypothesis: <i>Trade secrets were stolen.</i></p> <p>Entailment: FALSE</p>
<p>Premise: <i>It has been observed that in those countries of the world where capital punishment is still in operation, the crime rate, especially murder, is distinctively low in comparison to countries where capital punishment has been discarded.</i></p> <p>Hypothesis: <i>Capital punishment is a deterrent to crime.</i></p> <p>Entailment: TRUE</p>
<p>Premise: <i>A farmer who was in contact with cows suffering from BSE – the so-called mad cow disease – has died from what is regarded as the human form of the disease.</i></p> <p>Hypothesis: <i>Bovine spongiform encephalopathy is another name for the "mad cow disease"</i></p> <p>Entailment: TRUE</p>
<p>Premise: <i>The girl was found in Drummondville.</i></p> <p>Hypothesis: <i>Drummondville contains the girl.</i></p> <p>Entailment: FALSE</p>
<p>Premise: <i>The official visit of the Argentine minister marks a further step in the normalisation of UK-Argentine relations.</i></p> <p>Hypothesis: <i>Relations between Argentina and Great Britain are growing more cooperative.</i></p> <p>Entailment: FALSE</p>

Table 6: Some problematic instances we have found in the Recognizing Textual Entailment (RTE) dataset.

<p>Context 1: <i>I tried to make a call, but the <u>line</u> was dead.</i></p> <p>Context 2: <i>A dedicated <u>line</u>.</i></p> <p>Sense Match: TRUE</p>
<p>Context 1: <i>The author gives a depressing picture of life in Poland.</i></p> <p>Context 2: <i>He had no clear <u>picture</u> of himself or his world.</i></p> <p>Sense Match: FALSE</p>
<p>Context 1: <i>Instant replay caused too long a <u>delay</u>.</i></p> <p>Context 2: <i>The <u>delay</u> before the echo of a sound.</i></p> <p>Sense Match: FALSE</p>
<p>Context 1: <i><u>Stop</u> a car.</i></p> <p>Context 2: <i><u>Stop</u> the thief.</i></p> <p>Sense Match: TRUE</p>
<p>Context 1: <i><u>Fall</u> asleep.</i></p> <p>Context 2: <i>She <u>fell</u> to pieces after she lost her work.</i></p> <p>Sense Match: TRUE</p>

Table 7: Additional problematic instances we have found in the Word-in-Context (WiC) dataset.

the first answer. Finally, “80” is labeled as FALSE further contradicting the penultimate answer.

RTE In the RTE example (Table 2), the specific premise regarding *Pacific countries* is not sufficient to entail the general hypothesis, thus the answer should be FALSE. We provide more examples in Table 6. In particular, in some of them, we believe that the label is incorrect (examples 1, 3 and 5), or at least highly misleading, while in some others we think that not enough information is provided to entail the hypothesis (examples 2 and 4).

WiC In the WiC example provided in Table 2, the word *criticism* is used with the same meaning in the two contexts, namely *disapproval expressed by pointing out faults or shortcomings* according to WordNet. We provide additional ambiguous or wrongly annotated examples in Table 7.

By inspecting the WiC dataset, it is immediately apparent that, in many negative examples, the semantic gap between the meanings of the same lemma in the two contexts is very narrow. Although such cases are difficult even for machines, we posit that for humans (especially if sense distinctions are not provided and annotators are not lexicographers, as in WiC) they are way more difficult.

SQuAD For the SQuAD dataset, studies about errors in the annotations have already been performed by Rodriguez et al. (2021) through automatic error detection methods. Specifically, they annotated SQuAD items by discriminability, difficulty, and Item Response Theory (IRT) prediction errors, and discovered that items with negative discriminability, or where IRT’s prediction is wrong, have a much higher rate of annotation error, i.e. they are often “flawed” or “wrong”. We believe that tools for error detection (Klie et al., 2022) should play a key role in the improvement of existing benchmarks and in the creation of new ones.

Finally, still related to the topic of wrongly-annotated or ambiguous instances in the datasets, Nangia and Bowman (2019) performed an interesting study on the GLUE benchmark. In order to investigate the effect of these instances, they looked at human performance when there is 5-way annotator agreement. Using unanimous agreement has the effect of filtering out examples for which: i) the annotation guidelines supplied do not provide clear advice, and ii) humans understand the expectations of the task but find the example genuinely difficult or uncertain. They discovered that this widened the gap between humans and systems by more than 3 points on average, hence confirming the hypothesis that humans were often penalized by unclear guidelines or other factors. Even more interestingly, they found that in some tasks, when systems are evaluated on the unanimous subsets they obtain lower scores compared to those obtained on the entire test sets containing wrong or ambiguous instances, hence suggesting that systems had learned specific idiosyncrasies appearing in both training and test sets (Section 4.1).

Task	H	S	%
AX-b	100	1104	9.05%
AX-g	100	356	28.08%
BoolQ	100	3245	3.08%
CB	100	250	40.00%
COPA	100	500	20.00%
MultiRC	166	166	100.00%
RTE	500	3000	16.67%
ReCoRD	10000	10000	100.00%
WSC	146	146	100.00%
WiC	300	1400	21.42%

Table 8: Comparison of the size of the test sets annotated by humans (H) and systems (S) in the various tasks. % represents the ratio between H and S.

B Apples and Oranges

In Section 3.3, the first issue that we pointed out was that, on almost all SuperGLUE tasks, humans and machines are evaluated on different test sets (i.e. on a small subset vs. the full test set). Here, we provide more details (see Table 8). Specifically, it can be observed that only 3 out of 10 tasks are fully annotated by humans, while for the remaining 7 tasks only a small portion is annotated, ranging from 3% to 40% of the full dataset size.

C The Copenhagen Experiment

In this Section, we report on an experiment that was conducted in Copenhagen at the Danish Language Technology Conference in 2022⁸. The main goal was to verify the claim that contextual sentence examples from open lexical resources, such as those used to create the WiC dataset, i.e. WordNet, VerbNet and Wiktionary, “constitute a reliable base for the construction of the dataset, as they are curated in a way to be clearly distinguishable across different senses of the word” (Pilehvar and Camacho-Collados, 2019). Based on this assumption, “the [WiC dataset] annotators were not provided with knowledge from any external lexical resource”, and were asked to label each instance solely based on whether they thought the two occurrences of the word referred to the same meaning or not.

We repeated the above annotation task by asking 25 conference participants to provide “true” (T) and “false” (F) answers for the six instances in Table 9. The results show a high degree of disagreement, suggesting that the above claim is not always

⁸<https://cst.ku.dk/kalender/sprogteknologisk-konference-2022/>

WiC	Target	Context-1	Context-2	F	T
T	line (N)	I tried to make a call, but the <u>line</u> was dead.	A dedicated <u>line</u> .	14	11
F	love (N)	A mother’s <u>love</u> is not easily shaken.	The theater was her first <u>love</u> .	16	9
F	work (V)	This dough does not <u>work</u> easily.	<u>Work</u> the phones.	16	9
T	fall (V)	<u>Fall</u> asleep.	She <u>fell</u> to pieces after she lost her work.	17	7
F	picture (N)	The author gives a depressing <u>picture</u> of life in Poland.	He had no clear <u>picture</u> of himself or his world.	7	18
F	take (V)	Do you <u>take</u> sugar in your coffee?	A reading was <u>taken</u> of the earth’s tremors.	21	4

Table 9: Results of the Copenhagen experiment. Context-1 and Context-2 provide two sentences in which a Target word w appears. The WiC column specifies the label in the WiC dataset, indicating whether w has the same meaning in the two contexts. Finally, the F and T columns represent the number of people who voted for F and T, respectively.

valid, especially when subtle sense distinctions are involved. We posit that the presence of a certain amount of intrinsically debatable items hampers fair comparisons between humans and systems.

Indeed, in WSD evaluation tasks, where the granularity of senses is a key concern (Bevilacqua et al., 2021b), we advocate that the starting point for the task design should consist either of the warning made by many lexicographers that “there is very little agreement about what word senses are or how broad their scope should be, and no definitive way of knowing when one sense ends and another begins” (Atkins and Rundell, 2008), or the one from the famous lexicographer, James Murray, that “the best any lexicographer could hope for would be that readers would feel, on scanning a multisense dictionary entry, that this is not an unreasonable way of exhibiting the facts”. With large corpora and the latest advances in language modeling, we now have the possibility to measure differences between contexts in which words are used, and we need not and should not rely on made-up sentences from the times when corpora were not available at all. This is corroborated, for instance, by the inter-tagger agreement and the systems’ results of the multilingual version of the WiC task, where sentences come from real text and dictionary definitions are used as a help for annotators (Martelli et al., 2021).