No Strong Feelings One Way or Another: Re-operationalizing Neutrality in Natural Language Inference

Animesh Nighojkar, Antonio Laverghetta Jr., and **John Licato**
Advancing Machine and Human Reasoning (AMHR) Lab
University of South Florida
{anighojkar, alaverghett, licato} @usf.edu

Abstract

Natural Language Inference (NLI) has been a cornerstone task in evaluating language models' inferential reasoning capabilities. However, the standard three-way classification scheme used in NLI has well-known shortcomings in evaluating models' ability to capture the nuances of natural human reasoning. In this paper, we argue that the operationalization of the *neutral* label in current NLI datasets has low validity, is interpreted inconsistently, and that at least one important sense of neutrality is often ignored. We uncover the detrimental impact of these shortcomings, which in some cases leads to annotation datasets that actually *decrease* performance on downstream tasks. We compare approaches of handling annotator disagreement and identify flaws in a recent NLI dataset that designs an annotator study based on a problematic operationalization. Our findings highlight the need for a more refined evaluation framework for NLI, and we hope to spark further discussion and action in the NLP community.

1 Introduction

With the rise of large language models like GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and GPT-4,[1] it has become increasingly necessary to evaluate their language understanding and reasoning abilities. One influential task in this regard is natural language inference (NLI) (MacCartney and Manning, 2009, 2014), which is used to examine the inferential and commonsense reasoning skills of language models (Jeretic et al., 2020). NLI requires a model to determine the relationship between a statement, known as the *premise* P, and another statement, called the *hypothesis* H, by classifying it as *entailment* (H must be true given P), *contradiction* (H must be false given P), or *neutral* (H can or cannot be true given P).[2] NLI is crucial because it involves comprehending the logical properties of sentences, which is arguably a core capability of human reasoning and an important skill for language models to possess.

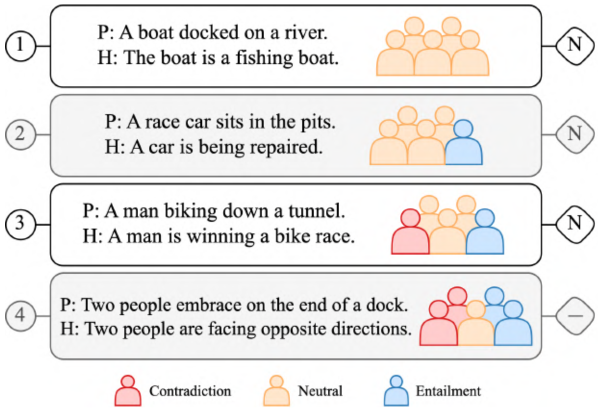


Figure 1: Selected NLI items from SNLI with annotations (shown by colors). The diamonds on the right show the gold label for these items in SNLI; note item 4 is marked '–' and is not assigned a gold label (hence it is ignored). We argue that items with all four annotation distributions should be considered neutral, but that there should be at least two sub-types of neutral.

Solving NLI requires the ability to perform textual inference between any two sentences (and in some cases, between any two arbitrarily long texts), making it a versatile framework for developing and evaluating reasoning benchmarks. Many NLP tasks, like question answering (Demszky et al., 2018), dialog systems (Gong et al., 2018), machine translation (Poliak et al., 2018), identifying biased or misleading statements (Nie et al., 2019), fake news detection (Yang et al., 2019), paraphrase detection (Nighojkar and Licato, 2021a,b), and fact verification (Thorne et al., 2018), require understanding and reasoning about the meaning of text and can be re-framed as NLI

[1] `https://openai.com/research/gpt-4`

[2] Recognizing textual entailment (RTE) (Dagan et al., 2006), a variant of NLI, only considers entailment and non-entailment.

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 199–210
July 13, 2023 ©2023 Association for Computational Linguistics

problems. NLI provides a broad framework for studying and alleviating logical inconsistencies in a language model's reasoning (Poliak, 2020; Mitchell et al., 2022) including explanation-based maieutic prompting (Jung et al., 2022), that uses NLI to evaluate individual links in a reasoning chain.

Most NLI datasets (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020a; Chen et al., 2020) utilize crowdsourcing to either generate NLI items or gather labels for pre-existing items. While this approach has advanced research on textual entailment, we believe that current NLI datasets, both established and recent, have overlooked important issues in their annotation design that hinder their validity as measures of textual entailment. Although the effects of different crowdsourcing schemes for NLI dataset development has been studied (Bowman and Dahl, 2021; Parrish et al., 2021), we focus on a specific issue: the operationalization of *neutral*. Neutral items usually have the lowest levels of annotator agreement (Nie et al., 2020b), and we contend that this disagreement has been handled improperly in previous work, contributing to the ongoing debate about how to handle disagreement in NLI (Palomaki et al., 2018; Pavlick and Kwiatkowski, 2019; Bowman et al., 2015; Williams et al., 2018). Instructions provided to annotators for labeling items as neutral are often ambiguous and inconsistent between datasets, with phrases like "neither" (Nie et al., 2020a) or "might be correct" (Bowman et al., 2015; Williams et al., 2018).

We believe these problems can be addressed by reconsidering the prevailing operationalization of neutral and replacing it with one which embraces disagreement. Although we are not the first to argue for the importance of properly incorporating disagreement (Palomaki et al., 2018; Pavlick and Kwiatkowski, 2019; Basile et al., 2021; Plank, 2022; Rottger et al., 2022; Uma et al., 2022b), we identify specific problems introduced by ignoring disagreement (for example, by dropping examples with low agreement entirely), and offer new evidence supporting its adoption grounded in the psychometric concept of *construct validity*. Consider the items shown in Figure 1, sourced from the SNLI dataset (Bowman et al., 2015). A general consensus on the gold label is reached by the annotators in the first three items, but the fourth item exhibits a high degree of disagreement. While the first three items are labeled neutral in SNLI and used to train models, the fourth is labeled with a special '-' class, indicating an irresolvable level of disagreement, and hence it is removed from training data (Bowman et al., 2015). This practice (also used by Williams et al.) effectively treats disagreement as an undesirable product of NLI data collection—a *linguistic annotation artifact* to be considered as noise rather than signal. But what is the source of this disagreement? Should item 4 in Figure 1 be ignored, or is it simply a different form of neutrality? We argue that item 4 should be considered a different sense of neutral than the one represented by item 1, because two interpretations are possible: (1) the individuals in the embrace may be facing in opposite directions, resembling a conventional embrace, and (2) one individual may be embracing the other from behind, thereby causing them to face the same direction. This ambiguity in how to interpret such items leads to two irreconcilable types of neutrals; items can be either *true* neutrals (item 1 in Figure 1), or they can be neutral as a result of *conflicting* interpretations (item 4).

Main contributions. In this paper, we address the aforementioned issues with neutrality in three ways:

1. We propose a new operationalization of neutral based on inter-annotator agreement, which we argue better captures two distinct senses of neutrality (true neutral and conflicting neutral) often conflated in NLI.

2. We compare our operationalization with a 4-way classification scheme based on annotator disagreement suggested by Jiang and de Marneffe (2019); Zhang and de Marneffe (2021); Jiang and de Marneffe (2022) and find that our operationalization has better construct validity, as using it to train models for NLI leads to better downstream performance.

3. We show that known limitations of at least one published NLI dataset (UNLI) are a direct consequence of its adopting an operationalization that did not embrace disagreement, instead opting to aggregate NLI annotations on a continuous scale. We analyze its methodological flaws, and make recommendations to avoid similar problems in future work.

2 Related Work

NLI is widely used for assessing language models' inferential capabilities, in part due to its generality and versatility. Many datasets, like SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), Adversarial NLI (ANLI) (Nie et al., 2020a), and WA-NLI (Liu et al., 2022) have been developed to evaluate a model's ability to reason through entailment relationships across a wide variety of contexts. Other datasets focus on specific domain knowledge (Holzenberger et al., 2020; Koreeda and Manning, 2021; Yin et al., 2021; Khan et al., 2022; Yang, 2022) or require knowledge of non-English languages (Conneau et al., 2018; Araujo et al., 2022).

In most NLI datasets, only one label per item is deemed correct, and models are tasked with determining the most plausible of three possible labels. However, there is a growing need for NLI tasks to handle a broader range of relationships and make finer-grained distinctions between them. Researchers are shifting their focus towards finer-grained annotations (Chen et al., 2020; Gantt et al., 2020; Meissner et al., 2021), as classical NLI tasks are not well-equipped to handle disagreement between annotators (Zhang et al., 2021; Zhang and de Marneffe, 2021; Jiang and de Marneffe, 2022; Wang et al., 2022). Recent research has also focused on assessing models' performance on *ambiguous* NLI items, where humans may disagree on the correct label. ChaosNLI (Nie et al., 2020b) was developed to study such ambiguities by gathering 100 human annotations on items from a subset of SNLI and MultiNLI, where only 3/5 of annotators agreed on the correct label. They found that models struggled to perform above random chance on items with low inter-annotator agreement and were unable to replicate the annotator label distribution (Zhou et al., 2022). Since most of the low agreement items are neutral (Nie et al., 2020b), we believe a possible reason for this poor performance is the conflation of true and conflicting neutrals as a single category (Section 4).

Zhou et al. (2022); Meissner et al. (2021) build on ChaosNLI and test language models' ability to recover the original annotator label distribution. However, the best results are still below estimated human performance. To solve ambiguous NLI items, Wang et al. (2022) argue that models need to be well-calibrated (i.e., their predicted probability distribution must correctly match the annotator distribution), and they show that label smoothing or temperature scaling can achieve competitive performance without direct training on the label distribution, though it should be noted that other work has found mixed success with using either of these approaches to address ambiguity in NLI (Uma et al., 2022a). According to Pavlick and Kwiatkowski (2019), annotator disagreements are *irresolvable* even when the number of annotators and context are both increased. Such items should not be ignored since the disagreement cannot be always attributed to noise. They argue that handling disagreements should be left to the ones using the models trained on these datasets. Similar to Zhou et al. (2022), Pavlick and Kwiatkowski (2019) also show that NLI models trained to predict one label cannot capture the human annotation distribution.

Despite calls in the literature for annotator disagreement to be accommodated rather than ignored, how this should be done has been the subject of much study. The earliest attempts from SNLI and MultiNLI simply assigned a '-' label to cases that had sufficiently low agreement, indicating that they should not be used for training (Bowman et al., 2015; Williams et al., 2018). More recent work has tried to incorporate low agreement items as a fourth *disagreement* class, a practice that began with Jiang and de Marneffe (2019) and was later used by Zhang and de Marneffe (2021); Jiang and de Marneffe (2022). We examine this practice in Section 3 and demonstrate that simply using a *catch-all* category for disagreement is not as effective as our operationalization for neutral items.

Another line of research has explored changing the annotation schema to use a continuous scale, rather than a discrete one, in the hope that this type of scale will better capture the subtleties of reasoning over ambiguity and lead to less disagreement. Chen et al. (2020) introduce *uncertain natural language inference* (UNLI), where annotators indicate the likelihood of a hypothesis being true given a premise. While models trained on UNLI can closely predict human estimations, later work has found that fine-tuning on UNLI can hurt downstream performance (Meissner et al., 2021), suggesting a serious flaw in the UNLI dataset. We analyze further issues with UNLI in Section 5.

In a recent study, Kalouli et al. (2023) propose a new interpretation of neutral based on the concept of *strictness*. They argue that, under "strict interpretation", the pair *P: The woman is cutting a tomato. H: The woman is slicing a tomato/* would be considered neutral as she could be cutting squares, but it could be considered an entailment pair if the interpretation is not so strict. Their operationalization of neutral based on the concept of *strictness* lacks clarity due to the absence of a precise, understandable definition of *strictness*. In effect, it simply shifts the problem of understanding what makes a pair of sentences neutral to understanding what makes their relationship "strictly logical" (a term they use to define strict interpretation, without further elaboration).[3]

3 Empirical evaluation of 'disagreement' as a fourth class

The classification scheme that uses a fourth 'disagreement' label for low-agreement items (Jiang and de Marneffe, 2019; Zhang and de Marneffe, 2021; Jiang and de Marneffe, 2022) conflates all three NLI labels in doing so. To explore this possibility, we conduct an empirical study to compare this disagreement-based scheme with other 4-way classification schemes. We define the *level of agreement* ($\mathbf{A}$) between annotators on NLI items as:

$$\mathbf{A} = \frac{\textit{number of votes for the majority label}}{\textit{total number of votes}} \tag{1}$$

We also explore two agreement threshold t values (0.8, and 1),[4] which is the cutoff-value of $\mathbf{A}$ below which items are considered to have "low agreement." Note that Jiang and de Marneffe (2019) choose $t = 0.8$ but do not provide an explanation for choosing it. We train ALBERT-base (Lan et al., 2019), DistilBERT-base-uncased (Sanh et al., 2019), Electra-base (Clark et al., 2020), DeBERTa-v3-base (He et al., 2020), and RoBERTa-base (Liu et al., 2019) to show that these results are not specific to just a few models. We are limited to using SNLI and MultiNLI because they are the only NLI datasets that report individual annotations in sufficient quantity to finetune transformer language models. We trained each model for 5 epochs and tested their performance on a held out, stratified, evaluation set.[5]. We use only the base versions of these models because our objective here is not to train the best models, but to examine and compare classification schemes. Models are being used in this experiment only to compare the *separability* of all classes for each of these classification schemes:

- **Con:** Entailment, Neutral, ↑ Contradiction, ↓ Contradiction [6]
- **Dis:** Entailment, Neutral, Contradiction, Disagreement
- **Ent:** ↑ Entailment, ↓ Entailment, Neutral, Contradiction
- **Neu:** Entailment, ↑ Neutral, ↓ Neutral, Contradiction

Better F_1 scores would suggest the model could better differentiate between the classes of the given classification scheme, and thus the scheme has better *ecological validity*.[7]

Results are shown in Figure 2. We find that using a fourth 'disagreement' label leads to the worst results consistently. These results suggest that having a catch-all 'disagreement' label does not provide enough information to help models successfully reason over ambiguous items. Note that unlike the other three schemes, **Dis** classifies all low-agreement items as 'disagreement', thus making the other three schemes more imbalanced than **Dis**. For instance, **Con** classifies only low-agreement contradiction items as the fourth class and low-agreement neutral and entailment items are classified as their respective majority labels. Lowest F_1 score on **Dis** (the most balanced classification scheme) is perhaps even more informative than it would have been if the schemes were equally balanced. Any of the other three schemes consistently leads to better

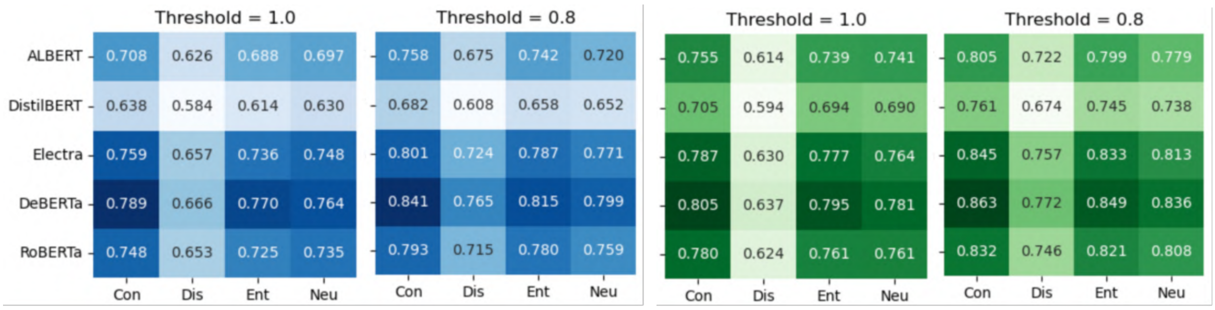
[3] Note that the strict conditional $\Box(p \rightarrow h)$ was famously introduced by Lewis (1912) as a formalization of the indicative conditional. However, this does not appear to be the sense of "strict" meant by Kalouli et al. (2023).

[4] Because SNLI and MultiNLI have at most 5 annotations, and the majority label is always taken as the gold label, 0.4 is the smallest possible $\mathbf{A}$ that can be used. Since all items at that agreement are marked as - in both the datasets, $t = 0.6$ cannot be used for **Ent** and **Con**. Also, $t = 0.6$ will give us same items for all four classes in **Dis** as well as **Neu**, making their comparison at that threshold meaningless.

[5] Github code will be released upon publication.

[6] ↑ and ↓ denote high and low annotator agreement respectively.

[7] Ecological validity examines whether the results of a study can be generalized to real-life settings (Egger et al., 2008).



(a) MultiNLI (b) SNLI

Figure 2: Heatmaps of F_1 scores on different 4-way classification schemes (x-axis) for different language models (y-axis). Darker boxes indicate better performance. Models consistently under-perform on the disagreement-based classification scheme (**Dis**) proposed by Jiang and de Marneffe (2019); Zhang and de Marneffe (2021); Jiang and de Marneffe (2022), indicating that a catch-all disagreement label does not provide enough information to models to reason over ambiguous items.

performance, regardless of model or threshold used, and thus has better construct validity (Bleidorn and Hopwood, 2019; Zhai et al., 2021) than the classification scheme based on disagreement.

4 Operationalizing Neutral

In NLI, the neutral label is used for situations where the relationship between the premise and hypothesis is ambiguous or there is insufficient information to determine the relationship. Neutral is often considered a catch-all for relationships that do not fall under entailment or contradiction. The definition of neutral is typically provided to crowd-source workers as “neither” (Nie et al., 2020a) or “might be correct” (Bowman et al., 2015; Williams et al., 2018).

But is a classification of neutral simply a default assumption that always means neither entailment nor contradiction can be definitively determined, or can it be a positive claim that a different type of relationship holds between the sentences? A closer look at the data obtained from NLI datasets suggests that neutrality is more complex than it may initially seem. According to Nie et al. (2020b), neutral items in many NLI datasets exhibit the lowest agreement levels. The most frequent label below an agreement level of $\mathbf{A} = 0.8$ for both the SNLI and MultiNLI subsets is neutral, while it is the least frequent label at a perfect agreement level. This lack of agreement motivates our focus on neutral particularly, as it is consistently the most problematic label to annotate. The empirical study in Section 3 also shows that a neutral-based classification scheme has a better separability than a disagreement-based classification scheme.

There are at least two senses in which the relationship between two sentences can be said to be neutral, which become clear if we imagine two possible justifications that an individual NLI annotator may provide for why they selected the label neutral: (1) *True Neutral:* The annotator cannot find any sufficiently strong reasons (using whichever standard of strength they determine appropriate) to satisfy either entailment or contradiction; or (2) *Conflicting Neutral:* The annotator finds strong reasons to support *both* entailment and contradiction.

It is a central position of this paper that these two interpretations of the neutral label are irreconcilable and should not be confused with each other. Attempting to conflate the two, e.g. by assuming that neutrality is simply the mid-point on a continuous scale between the two extremes of entailment and contradiction, will and has led to significant reductions in quality of data collections and their resulting benchmark datasets (see §5).

No existing NLI dataset, to our knowledge, asks or encourages annotators to explain whether their reasons for selecting neutral are in line with true or conflicting neutral as we have defined them above. For the present work, then, we present evidence for the discriminant validity of true and conflicting neutral (i.e., that they refer to two distinct constructs that can and should be measured separately Campbell and Fiske (1959)) by assuming that they will be *approximately reflected in the distribution* of individual annotations on a single NLI item—in other words, conflicting neutral items will tend to have annotation distributions resembling item

Dataset	Mean Length (T)	Mean Length (C)	Reading Ease (T)	Reading Ease (C)
$*$ SNLI dev + test	109.6	118.2	84.0	82.8
SNLI train	102.8	111.3	84.8	83.6
$*$ MultiNLI matched + mismatched	172.0	183.0	67.0	65.2
MultiNLI train	163.8	186.0	68.7	64.4
ANLI R3 dev	389.0	372.7	67.9	65.3
ANLI R3 test	382.4	392.7	69.8	66.1
ANLI R3 train	369.3	377.3	66.3	64.6
WA-NLI test	147.3	147.6	77.4	77.4
WA-NLI train	147.5	148.6	77.1	77.0

Table 1: Comparison of true (T) and conflicting (C) neutrals. Smaller values for reading ease indicate harder-to-read items. We use our trained model to estimate $\mathbf{A}$ for the datasets that do not release individual annotations and the ones that do are marked with a "$*$". Cases where our hypothesis was NOT confirmed are underlined and in brown.

4 in Figure 1, whereas true neutrals will tend to match item 1. Results in Section 3 show that indeed such a classification scheme does a much better job of separating the four classes for models than a scheme that conflates all three labels.

True vs. Conflicting Neutral: Surface-level Differences We perform an exploratory analysis to identify potential reasons why annotators may disagree on some 'neutral' items, to better motivate our operationalization of 'neutral'. Drawing from Pavlick and Kwiatkowski (2019), who found that disagreement increases as more context is given, we investigate whether ambiguity in NLI items arises due to increased complexity, leading to difficulties in accurately interpreting them. We measure this complexity using two metrics: mean length of the item in terms of number of characters (after the premise and hypothesis are joined with a space), and Flesch Reading Ease (Flesch, 1948), a commonly-used measure of text readability. Our findings, shown in Table 1, reveal that true neutral items are shorter and easier to read than conflicting neutral items. However, the observed difference in complexity between the two forms of neutrals is marginal and inconclusive. These results suggest that at least superficial qualitative differences exist between different types of neutrals, but more extensive research is needed to clarify the extent of these differences.

5 An Analysis of UNLI

We have argued that a carefully grounded operationalization of the neutral label is crucial for ensuring the reliability (performance should be free from random error) and validity of NLI. To demonstrate the issues that can arise if this caution is not taken, we next analyze a recent NLI dataset — Uncertain NLI (UNLI) (Chen et al., 2020). The UNLI dataset, when used for fine-tuning, appears to actually harm downstream performance (Meissner et al., 2021; Zhou et al., 2022; Wang et al., 2022). UNLI attempts to enhance NLI by converting the categorical labels for some SNLI items to a continuous scale. Participants were instructed to rate the likelihood of a given hypothesis being entailed by a given premise using an ungraduated slider, ranging from 0 (labeled as "impossible") to 1 (labeled as "very likely") and were shown the probability they were assigning to the *premise-hypothesis* pair in real time.

According to Chen et al. (2020), the probabilistic nature of NLI (Glickman et al., 2005) suggests that not all contradictions or entailments are equally strong.[8] Thus, UNLI was developed with the intention of capturing subtler distinctions in *entailment strength* using a continuous scale. This dataset has over 60K items from SNLI, annotated by humans. For each premise-hypothesis pair, two annotations were collected, and in cases where the first two annotators differed by 20% or more, a third annotator was consulted. However, the dataset only reports the averaged scores, which makes it impossible to assess the degree of agreement or correlation between the two annotators or even identify examples where a third annotator was needed. Thus, reported values near 0.5 (which we might take to be the equivalent of *neutral* items) fundamentally conflate items where both annotators chose the midpoint on the slider with items where each annotator chose one of the extremities.

The assumption that one continuous scale can capture even the three categories in standard NLI (entailment, contradiction, and neutral) is

[8]The view that NLI is inherently probabilistic, or that natural inference can be best modeled with probability, is not universally held, e.g. (Bringsjord, 2008).

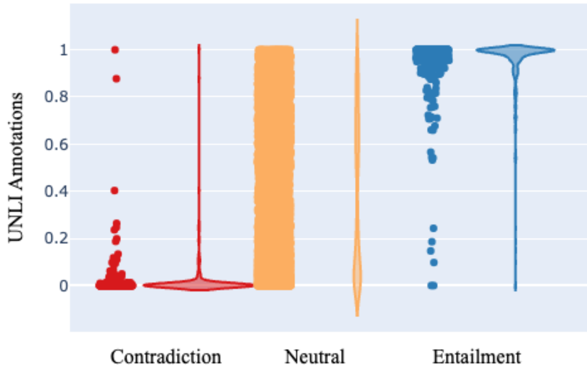


Figure 3: Figure 1 from Chen et al. (2020) redrawn on a linear scale. Note the two distinct bulges in the violin plot for neutral items, suggesting that annotators were confused about whether neutral items should be placed near 0 or middle of the slider.

a strong one (already shown to be problematic in (Pavlick and Kwiatkowski, 2019)), which is typically glossed over by presuming that entailment lies at the higher end of the spectrum, contradiction at the other end, and neutral somewhere in the middle. But no such instruction to interpret the scale this way was provided to annotators. Indeed, as we will show, annotators appeared to be confused as to whether an absence of entailment meant that the slider should be at the '0' position, or in the middle.

In their attempt to obtain subjective probabilities for premise-hypothesis pairs, the authors used a scale with 10K steps with a scaled logistic transformation ($f(x) = \sigma(\beta(x - 5000))$) to convert the values on the scale into probabilities between 0 and 1. They do not report the chosen value of β and do not specify whether the scores were averaged before or after applying the function, which is crucial information as both would yield different results. Because raw values of x are not provided, and we do not know whether scaling is performed before or after averaging, we are unable to recover the chosen values of β.

The scale Chen et al. (2020) used was based on EASL (Sakaguchi and Van Durme, 2018), an approach developed to collect scalar ratings in NLP tasks.[9] They then modified the EASL scale by utilizing the aforementioned logistic transformation, which they argued would allow for more nuanced values near both extremes. Notably, the source of the anchor points used on the scale (i.e., "impossible" and "very likely") is not explicitly stated by Chen et al. (2020), although it is possible they were obtained from JOCI (Zhang et al., 2017), a dataset created for studying ordinal commonsense reasoning that uses the same anchor points for opposite ends of the scale.[10]

In effect, their logistic transformation compresses the extreme ends of the scale, so that the graphic they display (Figure 1 in Chen et al. (2020)), at first glance, appears as if the NLI items labeled as contradiction, neutral, and entailment occupy roughly equal space across the continuum of values. Figure 3 instead depicts the distribution of averaged human responses collected by Chen et al. (2020) on a linear scale.[11] It is clear to observe in Figure 3 that while entailment and contradiction annotations are distinctly separated and skewed heavily towards the extreme opposite ends of the scale, annotations for neutral *span the entire range from 0 to 1*. The origin of this discrepancy is unclear, but based on the instructions given to them, it may be that annotators were unsure where to place neutral on the scale. Supporting this hypothesis is the bulge near 0 on the violin plot for neutral in Figure 3, which suggests that annotators chose 0 for both neutral and contradiction items. This information is obscured by the logistically transformed graph displayed by Chen et al. (2020).

Table 2 highlights some examples from UNLI that demonstrate the poor alignment of its annotations with SNLI annotation distributions. From Figure 3, the reliability of the scale for neutral annotations is notably poor, with annotations spanning the entire range of the scale. This suggests that neutral annotations lack internal consistency, an important measure of reliability (Rust and Golombok, 2014), because annotators do not label the NLI items in a consistent fashion even when the label remains constant.

Measurement issues are not uncommon in other fields that routinely run human studies, including psychological and educational mesurement. Development of annotation schemes in these fields often involves careful consideration of the item

[9]This scale was not validated for NLI by Sakaguchi and Van Durme (2018) and the tasks they evaluated it for — like evaluating quality of machine translations, or the frequency of words in language — differ significantly from NLI.

[10]This is further supported by the fact that Chen et al. (2020) cite Zhang et al. (2017) as a previous attempt to model likelihood judgments in NLI, which is also the aim of UNLI.

[11]Many of the properties of the scale we address here were unclear from reading the original figure in Chen et al. (2020), necessitating the redrawing.

Premise	Hypothesis	SNLI Annotations	UNLI score
A woman with a blue jacket around her waist is sitting on the ledge of some stone ruins resting.	A man sits on a ledge.	$4C-0N-1E$	0.88
A lady is standing up holding a lamp that is turned on.	She is lighting a dark room.	$2C-2N-1E$	0.78
A singer wearing a leather jacket performs on stage with dramatic lighting behind him.	A singer is on American idol.	$1C-4N-0E$	0.01
A small boy wearing a blue shirt plays in the kiddie pool.	Boy cooling off during the summer.	$1C-4N-0E$	0.89

Table 2: Items from UNLI along with their individual annotations from SNLI.

format, including the rating scale, to ensure that it effectively measures the construct of interest (Bandalos, 2018). This can be achieved through qualitative analysis, such as cognitive interviews and focus groups, where items are administered to test takers and feedback is collected to ensure that the scale is understood and completed accurately, among other things (Miller et al., 2014). However, in the development of UNLI, Chen et al. (2020) did not report using such procedures. Moreover, common practices in measurement research were missing from UNLI, such as reporting how bad-faith responses were identified and filtered out, using attention-check items (except the qualifying test, whose results are not provided as part of the dataset), employing a sufficienlty large sample size of annotators, and providing individual annotations and relevant information about the annotators like their recruitment and compensation. These omissions make precise scientific replication impossible, and raise concerns about the validity of UNLI as a measure of (and benchmark for) NLI, while also providing a plausible explanation for why prior research yielded poor results when using UNLI for fine-tuning.

6 Conclusion

In this paper, we examined the operationalization of neutral in NLI datasets. Our analysis revealed that previous attempts to handle ambiguity in NLI based on neutrality have significant issues with their validity as annotation strategies for NLI. We proposed a new operationalization of neutral into *true neutral* and *conflicting neutral*. Although instances of these forms of neutral are present in most popular NLI datasets, they have been conflated into one neutral label, limiting our ability to measure ambiguity in NLI effectively. We showed that this approach of casting NLI to a 4-way classification task is better than the disagreement-based classification scheme used in previous work. We used UNLI as a case study to highlight measurement and annotation issues that should be avoided in the future.

Of the many factors that make science successful, two of the most important are the ability to make carefully designed measurements, and replicability. The first of these cannot be met when measurements of constructs are made in ways that significantly compromise their validity and reliability. And replicability is made impossible when papers are published in reputable venues reporting unclear collection details, having important parameter choices omitted, and with datasets reporting summary statistics in place of crucially important data. A significant roadblock of the work we reported in this paper was the lack of availability of individual annotations in widely-adopted NLI benchmarks, even when there seems to be no public benefit in leaving out such information. It is our hope that the present work will encourage our fellow AI researchers to more highly value such considerations.

Limitations

We approximated the operationalization of the two senses of neutrality using annotator agreement. Perhaps a better basis for operationalizing the two senses of neutrality could be found in the reasons behind the annotators choosing the neutral label. Since no NLI datasets ask annotators to explain their choice and release those responses, we will try to analyze this in the future.

We presented a surface-level syntactic analysis of the differences between the two types of neutrals, but semantic differences should also be analyzed. Intuitively, semantic differences might give us a better understanding of these two types, but further study is needed to verify this.

Though we focus on UNLI as a case study to back up our claims, further analysis on a broader range of NLI datasets (and possible extensions to tasks beyond NLI) should also be conducted.

References

Vladimir Araujo, Andrés Carvallo, Souvik Kundu, José Cañete, Marcelo Mendoza, Robert E. Mercer, Felipe Bravo-Marquez, Marie-Francine Moens, and Alvaro Soto. 2022. Evaluation benchmarks for Spanish sentence representations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6024–6034, Marseille, France. European Language Resources Association.

Deborah L. Bandalos. 2018. *Measurement theory and applications for the social sciences*. Methodology in the Social Sciences. The Guilford Press, New York, New York ;.

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.

Wiebke Bleidorn and Christopher James Hopwood. 2019. Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2):190–203.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

Selmer Bringsjord. 2008. The logicist manifesto: At long last let logic-based artificial intelligence become a field unto itself. *Journal of Applied Logic*, 6(4):502–525. The Philosophy of Computer Science.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arxiv:2204.02311*.

Kevin Clark, Thang Luong, Quoc V. Le, and Christopher Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Matthias Egger, George Davey Smith, and Douglas Altman. 2008. *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

William Gantt, Benjamin Kane, and Aaron Steven White. 2020. Natural language inference with mixed effects. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 81–87, Barcelona, Spain (Online). Association for Computational Linguistics.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. A probabilistic classification approach for lexical textual entailment. In *AAAI*, pages 1050–1055. Pittsburgh, PA.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. *arXiv preprint arXiv:1709.04348v2*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. 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 6086–6091, Hong Kong, China. Association for Computational Linguistics.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*.

Aikaterini-Lida Kalouli, Hai Hu, Alexander F Webb, Lawrence S Moss, and Valeria de Paiva. 2023. Curing the sick and other nli maladies. *Computational Linguistics*, page 1–45.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

C. I. Lewis. 1912. Implication and the algebra of logic. *Mind*, 21(84):522–531.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pages 140–156.

Bill MacCartney and Christopher D. Manning. 2014. *Natural Logic and Natural Language Inference*, pages 129–147. Springer Netherlands, Dordrecht.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI 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 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.

Kristen Miller, Valerie Chepp, Stephanie Willson, and Jose-Luis Padilla. 2014. *Cognitive interviewing methodology*. Wiley Series in Survey Methodology. Hoboken, New Jersey.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. 2022. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. 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.

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

Animesh Nighojkar and John Licato. 2021a. Improving paraphrase detection with the adversarial paraphrasing task. 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 7106–7116, Online. Association for Computational Linguistics.

Animesh Nighojkar and John Licato. 2021b. Mutual implication as a measure of textual equivalence. *The International FLAIRS Conference Proceedings*, 34.

Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing*.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R Bowman. 2021. Does putting a linguist in the loop improve nlu data collection? *arXiv preprint arXiv:2104.07179*.

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

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

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the evaluation of semantic phenomena in neural machine translation using natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 513–523, New Orleans, Louisiana. 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.

John Rust and Susan Golombok. 2014. *Modern psychometrics: The science of psychological assessment*. Routledge.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022a. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022b. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Yuxia Wang, Minghan Wang, Yimeng Chen, Shimin Tao, Jiaxin Guo, Chang Su, Min Zhang, and Hao Yang. 2022. Capture human disagreement distributions by calibrated networks for natural language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1524–1535, Dublin, Ireland. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fake news detection as natural language inference. *arXiv preprint arXiv:1907.07347*.

Zhanye Yang. 2022. Legalnli: natural language inference for legal compliance inspection. In *International Conference on Advanced Algorithms and Neural Networks (AANN 2022)*, volume 12285, pages 144–150. SPIE.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Xiaoming Zhai, Joseph Krajcik, and James W Pellegrino. 2021. On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology*, 30:298–312.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Capturing label distribution: A case study in nli. *arXiv preprint arXiv:2102.06859*.

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.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.