VARIERR NLI: Separating Annotation Error from Human Label Variation

Leon Weber-Genzel * **Siyao Peng** * **Marie-Catherine de Marneffe** **Barbara Plank**
MaiNLP & MCML, LMU Munich, Germany
FNRS, CENTAL, UCLouvain, Belgium
{siyao.peng,b.plank}@lmu.de marie-catherine.demarneffe@uclouvain.be

Abstract

Human label variation arises when annotators assign different labels to the same item for valid reasons, while annotation errors occur when labels are assigned for invalid reasons. These two issues are prevalent in NLP benchmarks, yet existing research has studied them in isolation. To the best of our knowledge, there exists no prior work that focuses on teasing apart error from signal, especially in cases where signal is beyond black-and-white. To fill this gap, we introduce a systematic methodology and a new dataset, VARIERR (variation versus error), focusing on the NLI task in English. We propose a 2-round annotation procedure with annotators explaining each label and subsequently judging the validity of label-explanation pairs. VARIERR contains 7,732 validity judgments on 1,933 explanations for 500 re-annotated MNLI items. We assess the effectiveness of various automatic error detection (AED) methods and GPTs in uncovering errors versus human label variation. We find that state-of-the-art AED methods significantly underperform GPTs and humans. While GPT-4 is the best system, it still falls short of human performance. Our methodology is applicable beyond NLI, offering fertile ground for future research on error versus plausible variation, which in turn can yield better and more trustworthy NLP systems.

1 Introduction

Labeled data plays a crucial role in modern machine learning (ML) (e.g., Mazumder et al., 2023). Data quality impacts ML performance and, in turn, user trust. It is therefore of vital importance to aim at high-quality consistently-labeled benchmark data (e.g., Bowman and Dahl, 2021). However, recent research revealed a notable presence of *annotation errors* in widely-used NLP benchmarks (Klie et al., 2023; Rücker and Akbik, 2023). Similar observations were made recently in computer vision (CV) (Northcutt et al., 2021; Vasudevan et al., 2022; Schmarje et al., 2024).

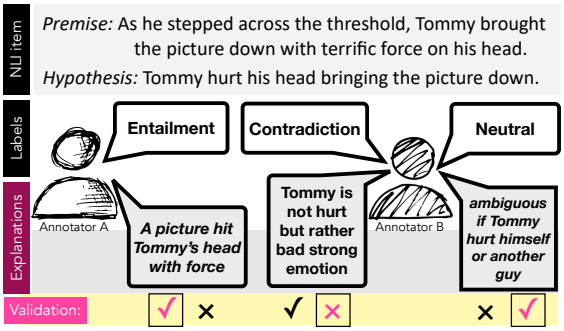


Figure 1: Variation or Error? We present a procedure and multi-label dataset, VARIERR, to tease apart annotation error from plausible human label variation. We leverage *ecologically valid explanations* and *validation* as two key mechanisms (boxed: self-validations; label "Contradiction" is an *error*); see §3-§4 for details.

At the same time, there is increasing evidence that for many items in many tasks, more than a single label is valid. For some items, systematic variation exists for valid reasons, such as plausible disagreement or multiple interpretations. In other words, the world is not just black and white. Human label variation (HLV, as termed by Plank 2022) has been shown on a wide range of NLP tasks (de Marneffe et al., 2012; Plank et al., 2014; Aroyo and Welty, 2015), including in natural language inference (NLI; Pavlick and Kwiatkowski 2019; Zhang and de Marneffe 2021). NLI involves determining whether a hypothesis is true (Entailment), false (Contradiction), or neither (Neutral), assuming the truth of a given premise; see Figure 1 for an example with plausible labels.

Although high-quality, consistently labeled data may initially appear to conflict with the goal of accommodating HLV, it is important to emphasize that we do not perceive these as contradictory goals. While HLV exists, so do errors. We assert that annotators are inevitably prone to make errors, such as misunderstanding instructions or accidentally

* Equal contribution.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2256–2269

selecting a wrong label. Optimizing data quality is essential through providing clear instructions and effective training, and identifying annotation errors yields better datasets (Larson et al., 2019). However, still little is known about what constitutes an error versus plausible variation. We lack both a theory and operationalizable procedures to tease apart error from plausible HLV consistently and soundly. Some datasets with errors (and their corrections) exist, and there has been work on automatic error detection (AED). However, both have their limitations (§2). A crucial gap remains: a lack of examination in real-world scenarios where the signal is nuanced, not merely black-and-white.

To address this gap, this paper contributes: (i) VARIERR, a novel multi-annotator English NLI dataset with both plausible variation and detected errors. To the best of our knowledge, no such dataset exists yet. (ii) A new methodology to detect errors: we collect multiple annotations, where each label comes with an ecologically valid explanation inspired by Jiang et al. (2023), and propose to pair them with validity judgments to identify errors. (iii) Finally, we benchmark existing AED methods and GPTs in a challenging setup, where the task is to tease apart error from plausible human label variation. Our findings indicate that existing AED methods underperform humans and GPTs substantially on our self-validated VARIERR NLI dataset. We release our data and code to facilitate uptake.[1]

2 Related Work

Labeled data is the fuel of machine learning, as it drives both learning and evaluation. Following a data-centric view, we focus on improving data quality over data quantity (Motamedi et al., 2021; Swayamdipta et al., 2020; Zhang et al., 2021; Gordon et al., 2022). We aim to bring together work on data quality from two ends: annotation error vs. human label variation.

Annotation Errors and AED Several recent work found errors in widely used benchmarks, such as CoNLL 2003 for Named Entity Recognition (Wang et al., 2019; Reiss et al., 2020; Rücker and Akbik, 2023), TACRED for relation extraction (Alt et al., 2020), WSJ for syntax (Manning, 2011; Dickinson and Meurers, 2003), and ImageNet for object classification (Beyer et al., 2020; Northcutt et al., 2021; Vasudevan et al., 2022).

AED has a long-standing tradition in NLP. Proposed methods range from early work that relies on variation-based methods positing that instances with similar surface forms tend to have the same label (Dickinson and Meurers, 2003; Plank et al., 2014) to more recent model-based approaches that either exploit predictions (Amiri et al., 2018; Arazo et al., 2019) or information derived from training dynamics (Swayamdipta et al., 2020); see Klie et al. (2023) for a survey on AED.

Flaggers and scorers for AED have been proposed (Klie et al., 2023). Flaggers detect errors by providing a hard decision on whether an instance is erroneous. Scorers, on the other hand, assign a score to each instance reflecting the likelihood of being an error, and the top-n scoring instances are then corrected. Here, we focus on scoring methods to rank instances. Most of the AED work mentioned has limitations as they either rely on post-hoc mining of errors (and might therefore miss out on errors) in semi-automatic ways (e.g., Reiss et al., 2020), or they inject synthetic noise which has been shown to result in datasets where errors are easy to spot (Larson et al., 2019). Instead of using synthetic noise, we focus on realistic setups and re-annotate data in ecologically valid ways.

Human Label Variation (HLV) Recent studies have drawn attention to HLV in NLP (i.a., Uma et al., 2021; Plank, 2022). HLV has been described as annotator disagreement, which is not just noise but also signal since a sign of vagueness or ambiguity can benefit models (Aroyo and Welty, 2013). These include judgments that are not always categorical (de Marneffe et al., 2012), inherent disagreement (Pavlick and Kwiatkowski, 2019; Davani et al., 2022), or justified and informative disagreement (Sommerauer et al., 2020). For subjective NLP tasks, which by essence encourage annotator subjectivity (and hence variation), there is also a line of work referred to as perspectivism (Cabitza et al., 2023), connected to the descriptive data annotation framework proposed by Rottger et al. (2022).

HLV in NLI This paper focuses on NLI, known to contain HLV (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang and de Marneffe, 2022; Jiang et al., 2023). Pavlick and Kwiatkowski (2019) re-annotated nearly 500 NLI instances with 50 crowdworkers and showed that disagreements in NLI cannot be dismissed as annotation "noise." ChaosNLI (Nie et al., 2020) pioneers large-scale NLI annota-

[1] https://github.com/mainlp/VariErr-NLI

tion by collecting 100 annotations per instance for 3K items from SNLI (Bowman et al., 2015), αNLI (Bhagavatula et al., 2020), and MNLI (Williams et al., 2018) but for which the original annotations did not yield high agreement. They show that, for most of the items, HLV persists with more annotations. Further, their experiments show a large room for model improvement and a positive correlation between human agreement and label accuracy.

In another line of work, Jiang and de Marneffe (2022) identified *reasons* for observing variation in NLI, deriving a taxonomy based on linguistic properties of the items. Following up on that work, Jiang et al. (2023) proposed LIVENLI, to gain insights into the origins of label variation. They re-annotated 122 NLI instances from ChaosNLI with *ecologically valid explanations*: annotators are instructed to not only provide NLI labels but also explanations for their label choices. This addresses a limitation of prior work that uses post-hoc explanations, which may not reflect the true reasons of the original annotators, thereby questioning the validity of the prior method. They show that ecologically valid explanations have an additional benefit: signaling *within-label variation*, i.e., annotators give the same label but for different reasons. While we do not focus on the latter here, we take inspiration from Jiang et al. (2023) to collect ecologically valid explanations (cf. §3.1).

To the best of our knowledge, there remains a gap for studies on *both* annotation errors and human label variation in a concentrated effort. It is thus an open challenge to define error in an ecologically valid way, and it is unknown to what extent existing AED methods help detect such errors and whether new methods are needed. To find answers to these challenging open questions, we believe it is important to move both directions forward.

3 VARIERR: Annotation Procedure

To tease apart human label variation from error, we create VARIERR (Variation versus Error), a NLI dataset with two rounds of annotations by four annotators:[2] Round 1 for NLI labels and explanations (§3.1) and Round 2 for validity judgments (§3.2). Table 1 presents two VARIERR examples with two-round annotations, as well as their deduced label variations and errors to be discussed in §4.

[2]Annotators are Master's students in Computational Linguistics and the first author of this paper, all paid according to national standards.

3.1 Round 1: NLI Labels & Explanations

We collect annotations from four annotators on 500 NLI items randomly sampled from the MNLI subset of ChaosNLI. Annotators were asked to provide not only one or more NLI labels (E: Entailment, N: Neutral, C: Contradiction) to each item but also a one-sentence *explanation* for each label they chose, as the same label could be chosen for different reasons (Jiang et al., 2023). Annotators could use a fourth "I don't know" (IDK) label if none of the NLI labels seemed suitable. The Round 1 annotation sums up to 1,933 label-explanation pairs with the standard three NLI labels for the 500 items, and 331 "IDK" annotations (released in the data) which are discarded in Round 2.

3.2 Round 2: Validity Judgments

VARIERR's key contribution lies in proposing a second round of *validity judgment*. Validity judgment mirrors conventional annotation adjudication in that annotators judge each other's NLI labels and explanations. This information is delivered anonymously to annotators to reduce group dynamics. However, rather than agreeing on a single label or explanation altogether, annotators are free to make independent judgments on each label-explanation pair from Round 1, which enables inferring what is an error versus plausible variation (cf. §4.2).

So in Round 2, annotators become judges. For all 500 items, the 1,933 label-explanation pairs from Round 1 are distributed anonymously to the same four annotators. For each NLI item, each judge sees all label-explanation pairs annotated in Round 1, including their own, which they may or may not remember.[3] For each label-explanation pair, the annotator judges whether the explanation makes sense for the NLI label, answering "yes" (✓), "no" (✗) or "IDK" (**?**, I don't know) as shown in the four right columns of Table 1. Round 2 amounts to 7,732 validity judgments, including 158 "IDK"s.

4 VARIERR: Detecting Errors

Multiple validity judgments on label-explanations enable distinguishing annotation errors from HLV.

[3]Round 1 took place August-November 2023, and each annotator spent ~26 hours to annotate 500 items. Round 2 took place from November 2023 to January 2024, and each spent ~15 hours. Annotators worked independently in different weeks. The interval between the end of Round 1 and the start of Round 2 was one month or longer for all annotators.

Premise: They made little effort, despite the Jesuit presence in Asia, to convert local inhabitants to Christianity or to expand their territory into the interior.
Hypothesis: The Jesuit presence in Asia helped to convert local residents to Christianity, allowing them to expand their territory.
Label-explanation pairs: Before:{E:1,C:4} Self-validated:{C:3} Peer-validated:{C:4}
Label: [C] Error:[E]

		Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	A	Explanation	1	2	3	4
E	1	Both premise and hypothesis suggest that the speaker does not understand.	✗	✗	✗	✗
C	1	The Jesuit didn't make much effort to convert local residents to Christianity or to expand their territory.	✓	✓	✓	✓
	2	They did not try to expand their territory.	✓	?	✓	✓
	3	The Jesuit did not make effort to convert local residents to Christianity or to expand their territory.	✓	✓	✓	✓
	4	They made little effort to convert the locals or to expand their territory. So they did not help.	✓	✓	✓	✓

(a) *id:* 28306c

Premise: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.
Hypothesis: Newspaper preprints can cost as much as $5.
Label-explanation pairs: Before:{E:1,N:2,C:1} Self-validated:{N:2} Peer-validated:{N:2,C:1}
Label: [N] Errors:[E, C]

		Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	A	Explanation	1	2	3	4
E	4	5 dollars for a piece of newspaper.	✗	✗	✗	✗
N	1	The context only mentions how low the price may be, not how high it may be.	✓	✓	✓	✓
	3	The maximum cost of newspaper preprints is not given in the context.	✓	✓	✓	✓
C	2	The context says 5 or 6 cents, not $5.	✗	✗	✓	✓

(b) *id:* 72870c

Table 1: Sample VARIERR NLI annotations. L: Label, A: Annotator; E: Entailment, N: Neutral, C: Contradiction; ✓: 'yes (makes sense)'; ✗: 'no'; **?**: 'IDK (I don't know)'; magenta: self-judgments, black: peer-judgments, Err: label error. Curly brackets in *label-explanation pairs* denote label counters, e.g., in 1a, Before:{E:1,C:4} means that there are one entailment and four contradiction label-explanation pairs before validation.

4.1 Self versus Peer

One consequential feature of our two-round multi-annotator procedure is the post-hoc distinction between self- vs. peer-judgments. *Self-judgments* refer to Round 2 judgments on the judge's own Round 1 label-explanation annotations (✓ , ✗ , ? in Table 1), whereas *peer-judgments* refer to judgments from other annotators (✓, ✗, **?**). Since we have four annotators, each label-explanation pair receives one self-judgment and three peer-judgments. Note that the self vs. peer distinction only enters into effect after data collection.

4.2 Validating Labels

Let $\mathcal{A} = \{a_1, .., a_4\}$ be the set of annotators.

Self-validated Label-Explanation A label-explanation pair given by annotator a_k on an item in Round 1 is *self-validated* if a_k marks the label-explanation pair as "yes" in Round 2.

Peer-validated Label-Explanation A label-explanation pair given by annotator a_k in Round 1 is *peer-validated* if the majority (≥2) of the other three annotators $\mathcal{A}\backslash\{a_k\}$ marks the pair as "yes" in Round 2.

For example, the item in Table 1a received the Contradiction (C) label and accompanying explanations from all four annotators in Round 1. Among these four label-explanations, three are *self-validated* (✓) in Round 2. On the other hand, all four explanations for C are *peer-validated* since the majority (all in this case) of the peers voted "yes" (✓) for each label-explanation.

4.3 What counts as an error?

In the conventional setup of annotation adjudication, multiple annotators discuss the rationales for their labels and converge to an agreed label. The annotations that are originally different from and subsequently changed to the agreeing label are considered annotation errors. Similarly, in VARIERR, a label-explanation pair might be considered wrong in retrospect (i.e., in Round 2) by the annotator who wrote it after reading all label-explanation pairs given to that item by all annotators.

Thus, in this paper, we use Round 2 *self-*

Validation	FreqType	E	N	C	∑	IAA
before validation	*repeated*	554	977	402	1,933	0.35
	aggregated	263	403	212	878	
self-validated	*repeated*	467	916	329	1,712	0.50
	aggregated	210	380	159	749	
peer-validated	*repeated*	446	859	296	1,601	0.69
	aggregated	177	335	130	642	

Table 2: Frequency counts and inter-annotator agreement (Krippendorff's α with MASI-distance) on non-, self-, and peer-validated VARIERR NLI labels.

judgments—approving or rejecting annotators' own Round 1 label-explanation pairs—as the criteria for annotation errors.[4] We define a NLI label as an *error* if all label-explanation pairs are not self-validated. In other words, a label is viewed as correctly attributed to an item if any of its explanations is self-validated. In Table 1a, the Contradiction (C) label has at least one self-validated (✓) explanation (it even has three), and is thus not deemed an error. In contrast, Entailment (E) is an error in Table 1a because none of its explanations is self-validated, similarly for E and C in Table 1b.

4.4 Data Statistics & IAA

Table 2 shows the frequencies of NLI labels across the four annotators on the 500 items and 1,933 explanations before and after validation. We include statistics on *repeated* frequency counts (e.g., E counts twice if it is given as a label by two annotators for the same item) and *aggregated* labels (repeated labels for a given item count once). Moreover, following Jiang et al. (2023), we compute inter-annotator agreement (IAA) on NLI labels using Krippendorff's α (for multi-annotator) with MASI-distance (for multi-label).

Since all VARIERR items are sampled from ChaosNLI, which only includes MNLI items with two or all three of the NLI labels in the original annotations by design, we expect HLV and, thus, a medium-to-low IAA in our dataset. Indeed, VARIERR has an IAA of 0.35 (Krippendorff's α with MASI-distance) before validation, which raises to 0.50 and 0.69 after self- and peer-validations, with the latter reaching substantial agreement (see A.1 for pairwise IAA). However, with the appreciation of HLV, as long as there are no errors in the data, we argue that perfect agreement is not reachable. As a matter of fact, though the repeated and aggregated frequencies of NLI labels decrease adequately after validation, HLV still exists in self- and peer-validated annotations, averaging 1.50 (749/500) and 1.28 (642/500) labels/item.

We also observe in Table 2 that 88.57% (1,712/1,933) of Round 1 explanations in VARIERR were self-validated and 82.82% (1,601) were peer-validated. Figure 2a presents the number of label-explanation pairs rejected by both self- and peer-validations, by self-validation only, and by peer-validation only. Most Entailment and Contradiction annotations rejected by self are also rejected by peers (dark green). However, Neutral presents a challenging situation for self-validation where 60.13% (92/153) of Ns are only invalidated by the joint force of peers but not by one annotator alone.

Figure 2b gives the frequencies of aggregated label combinations per item before validation and after self- and peer-validations (see A.2 for label-explanation pair frequencies). Frequencies of multi-labeled items drop after self-validation and, more remarkably, after peer-validation. Inversely, the number of single-labeled items increases vastly, especially for Neutral. We also observe from VARIERR that a large portion of items, 37.6% (188/500), are self-identified as errors and 51.6% (258) are rejected by peer-validation.

In sum, though HLV remains in VARIERR, our validation process demonstrates that annotation errors are frequently concealed under label variations. We thus proceed with the challenging automatic error detection task in §5-6 to separate annotation errors from valid HLVs.

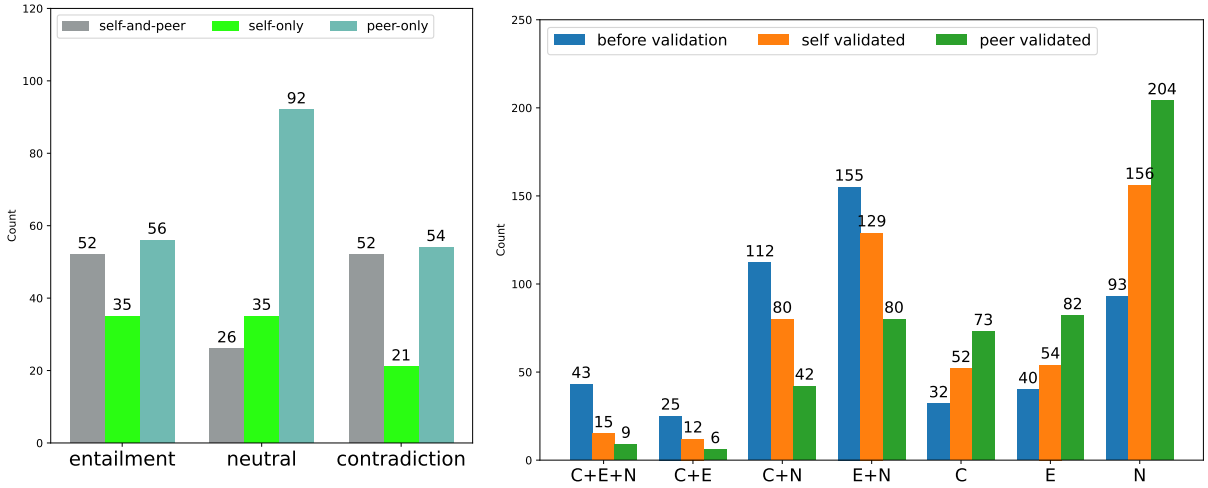
5 Automatic Error Detection (AED) on VARIERR

We now describe our experiments to detect annotation errors using VARIERR automatically. We evaluate the capabilities of AED methods, LLMs, and human heuristics (all henceforth *scorers*) in capturing annotation errors.

5.1 Task Definition and Evaluation

Following Klie et al. (2023) and Weber and Plank (2023), we model AED as a ranking task. In this setting, the goal of the *scorer* is to provide a ranked list with the labels that are most likely errors at the top and the most likely correct ones at the bottom. This ranked list can then be used to guide re-annotation efforts (Alt et al., 2020; Northcutt et al., 2021) or remove the most likely errors from the

[4] We opted for a strict error definition here. Peer validation could be used in the future but requires additional decisions.



(a) Number of label-explanations rejected by self-and-peer, self-only, and peer-only validations.

(b) NLI label sets on non-, self- and peer-validated items.

Figure 2: Frequency statistics on VARIERR.

training data (Huang et al., 2019). Scorers produce such a list by assigning an error score to each assigned label in the dataset. They derive the ranked list by sorting it based on the assigned scores.

We evaluate scorers on VARIERR using the following protocol. A model receives the list of NLI items from VARIERR where each item is paired with the *aggregated* label(s) it received in Round 1. For the 500 items in VARIERR, the model is given a list of 878 item-label pairs (cf. Table 2). Based on that information, the model assigns an error score and ranks the labels by this score. We evaluate how well the model performs by comparing this ranked list with the self-flagged errors (§4.3). Following Klie et al. (2023), we use standard ranking metrics for evaluation: average precision (AP), i.e., the area under the precision/recall curve computed over all assigned labels, and precision/recall for the top 100 ranked labels, P@100 and R@100.

5.2 Models

We evaluate five different AED models: two variants of Datamaps (DM, Swayamdipta et al. 2020), Metadata Archaeology (MA, Siddiqui et al. 2023), and two GPTs. We report the mean and standard deviation over three random seeds for DM and MA.

Datamaps (DM) We use training dynamics (i.e., the collection of training statistics over epochs E) for each label. These statistics are obtained by training a DistilRoBERTa-base model[5] (Sanh et al., 2019) following Klie et al. (2023) in a multi-label setting (Jiang and de Marneffe, 2022) on all labels of VARIERR obtained in Round 1. We refer to the j'th label of the i'th example as $\text{label}_{i,j}$. The training dynamics are modeled by the probability $p_{i,j,e}$ that DistilRoBERTa predicts for $\text{label}_{i,j}$ after the e'th epoch. Based on these probabilities, the two DM models we use are defined as follows:

$$DM_{\text{mean}} = -\frac{1}{E}\sum_{e=1}^{E} p_{i,j,e} \quad (1)$$

$$DM_{\text{std}} = \sqrt{\frac{1}{E}(\sum_{e=1}^{E} p_{i,j,e} + DM_{\text{mean}})^2} \quad (2)$$

Note that a *low* average probability for the label indicates a likely error. Because our evaluation setup requires the most likely errors to be ranked first, we negate the average probabilities.

Metadata Archaeology (MA) MA models AED as a supervised task. It represents each instance (or label in our case) as the E-dimensional $-\log p_{i,j,e}$ vector, where E is the number of epochs and $p_{i,j,e}$ is the probability the model assigns to the j'th label of the i'th NLI instance at epoch e. Then, it assumes that some instances are labeled with the property of interest (in our case, whether it is an erroneous label). It predicts whether an instance is an error by employing a k-nearest neighbors (kNN) classifier using the instance representations and error labels. We use the average number of annotated errors for the kNN to obtain a score for each instance. Following Siddiqui et al. (2023), we use $k = 20$. To obtain unbiased predictions, we require that the kNN training instances are distinct from those we want to obtain predictions for. We use a 2-fold cross-validation setup where we split VARIERR

[5] https://huggingface.co/distilroberta-base

into two folds, use one half as ground truth, obtain the predictions for the other, and vice versa.

GPT We also compare two large language models (LLMs): GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI, 2023).[6] We emulate the Round 2 annotation setting in §3.2 as closely as possible by prompting each model to provide a score reflecting how much each Round 1 explanation makes sense for a given label. We compute the score per label by averaging the GPT-assigned scores of all explanations for the label. We prompt GPT as follows, giving it the premise (context) and hypothesis (statement) of a NLI item as well as all label-explanation pairs, asking it then to provide a probability for each reason:

```
System:
You are an expert linguistic annotator.

User:
We have collected annotations for a NLI
     instance together with reasons for the
     labels. Your task is to judge whether the
     reasons make sense for the label. Provide
     the probability (0.0 - 1.0) that the
     reason makes sense for the label. Give
     ONLY the reason and the probability, no
     other words or explanation. For example:

Reason: <The verbatim copy of the reason>
Probability: <the probability between 0.0 and
     1.0 that the reason makes sense for the
     label, without any extra commentary
     whatsoever; just the probability!>.

Context: {CONTEXT}
Statement: {STATEMENT}

Reason for label {LABEL}: {REASON_1}
Reason for label {LABEL}: {REASON_2}
[...]
Reason for label {LABEL}: {REASON_n}

Reason {REASON_1}
Probability:
```

We implement GPTs using sglang (Zheng et al., 2023) and its default sampling parameters. See Appendix B for a complete prompt example. Note that the GPTs have access to the explanations for the labels, whereas the other models described above only have access to the labels without explanations.

5.3 Human Heuristics

In addition to the above automatic means, we experiment with four human heuristics that use the human label distributions over NLI labels (E, N, C) from annotation efforts: label counts from ChaosNLI (100 annotators) and VARIERR (4 annotators). In addition, we compare to VARIERR's total and average peer judgments over explanations.

Label Counts (LC): ChaosNLI & VARIERR We hypothesize that if multiple annotators choose the same label, there is a high likelihood that the label is a correct annotation. We implement two label count (LC) baselines: one using ChaosNLI (Nie et al., 2020) and one using VARIERR.[7] Since VARIERR is a subset of ChaosNLI items, we use label counts from ChaosNLI (LC_{CHAOS}) as a human heuristic to score Round 1 labels on each item, i.e., how many of the 100 crowd-workers annotated $label_{i,j}$ on item i. For instance, the ChaosNLI human distribution is {N:25, E:72, C:3} for the example in Figure 1. Similarly, we include $LC_{VARIERR}$ that counts the number of annotators (4 in total) that assigns $label_{i,j}$ to item i in VARIERR's Round 1 NLI labels. We multiply both LC_{CHAOS} and $LC_{VARIERR}$ by -1, proposing that if a label has a higher count, then it is less likely to be an error.

Peer-judgments (Peer) in VARIERR VARIERR's 2-round annotations enable more fine-grained human heuristics that engage judgments on label-explanation pairs. Since each $label_{i,j}$ can be assigned by multiple annotators with different explanations, we count the number of "yes" judgments on explanations from peers, i.e., excluding self-judgments since those are used for gold error labels.

We implement two peer heuristics: $Peer_{sum}$ and $Peer_{avg}$. $Peer_{sum}$ sums all "yes" judgments across multiple explanations on the same label, and $Peer_{avg}$ sums "yes" judgments within each explanation and then averages across explanations within the label. Given that one label can maximally receive four explanations, it can receive up to 12 peer-judgments, 3 per explanation. For example, C in Table 1a receives 11 peer-judged "yes" in total ($Peer_{sum} = 3+2+3+3 = 11$), and the average over four explanations is $Peer_{avg} = 11/4 = 2.75$. $Peer_{avg}$ differentiates more from $Peer_{sum}$ when there are multiple explanations, but each receives sparse "yes" judgments. For example, N in Table 5a (Appendix C) receives 3 "yes" judgments but across two explanations, resulting in $Peer_{sum} = 2+1 = 3$ and

[6] GPT models have previously seen the premise-hypothesis pairs from MNLI (Balloccu et al. 2024 and https://hitz-zentroa.github.io/lm-contamination/), but not the ample new annotations, i.e., NLI labels, explanations, and self/peer-validation judgments from multiple annotators.

[7] We did not include a comparison with LiveNLI (Jiang et al., 2023) because among its re-annotated 122 MNLI items, only 15 are shared with VARIERR.

$\text{Peer}_{\text{avg}} = 3/2 = 1.5$. Similarly to the label counts above, we multiply both Peer_{sum} and Peer_{avg} by -1, hypothesizing that fewer "yes" judgments indicate a higher likelihood to be an annotation error.

Combining Label Counts and Models Ranking labels by the number of annotations they received in Round 1 is a very strong baseline; see $\text{LC}_{\text{VariErr}}$ in Table 3. Inspired by Nogueira et al. (2019), we investigate an approach that re-ranks the predictions of $\text{LC}_{\text{VariErr}}$ by breaking ties with the scores produced by another model (e.g., DM, MA or GPTs). Note that $\text{LC}_{\text{VariErr}}$ produces many ties because its score is always one of $\{-1, -2, -3, -4\}$.

6 Results for AED on VariErr

Table 3 presents human and model performances on VariErr AED using the ranking setup in §5.

Scorer	AP	P@100	R@100	AP (rerank)
		Baselines		
Random	14.7	14.7	11.4	-
		Models		
MA	17.7 ± 1.5	18.3 ± 4.2	14.2 ± 3.2	44.2 ± 3.0
DM_{mean}	22.8 ± 0.4	23.7 ± 2.1	18.3 ± 1.6	**50.4** ± 0.7
DM_{std}	22.3 ± 1.9	22.7 ± 1.2	17.6 ± 0.9	50.0 ± 1.5
GPT-3.5	17.6	21.0	16.3	37.6
GPT-4	**31.3**	**46.0**	**35.9**	47.4
		Human		
LC_{Chaos}	32.5	35.0	27.3	**49.8**
$\text{LC}_{\text{VariErr}}$	40.8	42.0	32.6	40.8
Peer_{avg}	42.2	46.0	35.9	47.8
Peer_{sum}	**46.5**	**47.0**	**36.7**	47.8

Table 3: Results for AED on VariErr. AP: average precision; rerank denotes using the method to break ties in $\text{LC}_{\text{VariErr}}$. For MA and DM, we report mean and standard deviation over three random seeds. Note that GPTs have access to explanations.

6.1 Human Performance

The best human heuristic is from peers (Peer_{sum}), reaching a performance of 46.5% AP, 47% precision@100, and 36.7% recall@100. These numbers support our hypothesis that human validation can be used as a strong means to detect annotation errors in a task with relatively high HLV because self- and peer-rejected label-explanation pairs overlap considerably (cf. Figure 2a). Interestingly, both peer-derived heuristics from VariErr perform better than LC_{Chaos} (3 linguists versus 100 crowd-workers), which suggests that having few highly-trained expert annotators is sufficient for reliable error detection, outperforming a larger number of crowd-workers. $\text{LC}_{\text{VariErr}}$ outperforming LC_{Chaos} on all metrics strengthens this finding. Next we compare humans to automatic means.

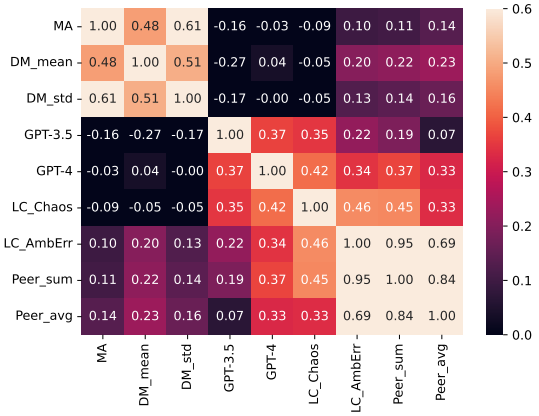


Figure 3: Correlations among scorer predictions.

6.2 Model Performance

Among the models, GPT-4 outperforms all other methods by a large margin with a 8.5/22.3/17.6 percentage points (pp.) improvement in terms of AP / P@100 / R@100 over the second best model DM_{mean}. GPT-4 even outperforms LC_{Chaos} in P@100 and R@100 and is close to the best peer heuristic for these two metrics.

One might postulate that ChaosNLI could have been part of GPT-4's training mixture, and GPT-4 performed well by reproducing its probabilities. To check whether this is the case, we compute Pearson's r between the predictions of all scorers (Figure 3). While GPT-4 has a slightly higher correlation (0.42) with LC_{Chaos} than with all other methods, it is still much lower than some correlations between other models, e.g., 0.61 between DM_{std} and MA. Thus, we conclude that GPT-4 does not solely rely on information from ChaosNLI but achieves its strong performance via some other mechanism. Another possible explanation is that it is the only model next to GPT-3.5 that has access to explanations. In the future, we would like to investigate the use of explanations further.

Moreover, Figure 3 allows for a more general interesting observation. There seems to be a clear cluster structure in which the training-dynamics-based models (DM and MA) correlate highly with each other and GPT-4 clusters with human scorers. Notably, correlations across these two clusters are small to non-existent or sometimes even negative.

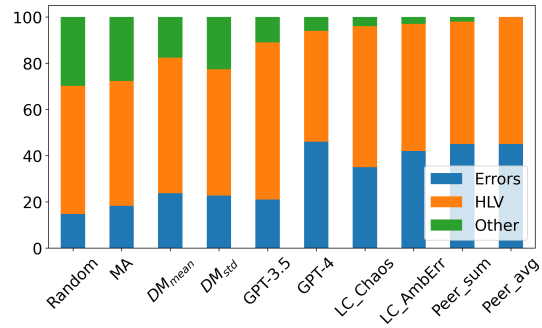


Figure 4: Average distribution of erroneous, HLV, and other labels over the top 100 instances per method.

6.3 Influence of Human Label Variation

In which situations do AED methods make mistakes, such as detecting false positive errors? This is an open question. We hypothesize that many top-ranking labels would either be errors or come from instances displaying HLV, i.e., instances with multiple labels after self-validation. The rest should be instances with just one plausible label. To test this hypothesis, we compute the proportion of *erroneous labels* vs. *valid labels from HLV instances* vs. *other* (with a single plausible label and thus exhibiting neither errors nor HLV labels) for the top 100 ranking labels for each method.

The results for the GPTs and human heuristics in Figure 4 confirm our hypothesis: they place very few (0-11) labels that are neither errors nor HLV in the top 100. On the other hand, the training-dynamics-based methods MA and DM assign between 17.6 and 29.8 of these items to the top 100. This suggests that increasing the separation between errors and HLV is only one part of improving training dynamics methods for AED. Another could be finding the characteristics of the top-ranking items that are neither errors nor HLV.

6.4 Reranking models using label counts

Column *AP (rerank)* in Table 3 presents our re-ranking results. We observe that re-ranking improves over vanilla LC$_{\text{VariErr}}$ for all methods but GPT-3.5. Interestingly, the best performing methods—also compared to the non-re-ranking approaches—are DM$_{\text{mean}}$ and DM$_{\text{std}}$. They even perform better than Peer$_{\text{sum}}$, the best human approach. This suggests that combining statistics from multiple annotators with AED methods based on training dynamics is a promising future direction.

7 Conclusion

Errors exist in datasets, but so does plausible human label variation. This paper defines a general procedure to separate the two by leveraging ecologically valid explanations (where annotators provide their reasons for a label) and pairing these with annotators' validations (to allow corrections). We provide a new VariErr dataset for the task of NLI re-annotated from scratch. Our empirical investigation on VariErr for NLI finds that traditional annotation error detection methods fare poorly and underperform humans and LLMs.

While this paper only applies our 2-round annotation procedure, VariErr, to NLI data, our methodology is general, and we hope it inspires uptake. Future work includes adapting these approaches to other NLP tasks, probing differences between self- and peer-judgments, mapping such strategies to (large) language models, and linking VariErr to experiments with LLMs' explainability, self-correction, or multi-agent systems.

Limitations

We believe that our two-round annotation setup would work for eliciting ecologically valid error annotations in tasks or languages other than English NLI. However, we cannot be sure without trying it, which we did not do in this project. Further, we did not use all types of information that VariErr contains for the training-dynamics-based AED methods. An interesting question would be whether exploiting the soft label distribution with methods from learning from disagreement (Uma et al., 2021) would improve AED results. Another potentially useful source of information is the explanations given by the annotators. Using this information for computing the training dynamics or directly modeling whether an explanation makes sense for a label in a supervised setting could potentially improve AED performance.

Acknowledgements

We thank Huangyan Shan, Shijia Zhou, and Zihang Sun for their contributions and invaluable feedback on VariErr. Thanks also to Verena Blaschke for giving feedback on earlier drafts of this paper, as well as to the reviewers for their feedback. Marie-Catherine de Marneffe is a Research Associate of the Fonds de la Recherche Scientifique – FNRS. This work is funded by ERC Consolidator Grant DIALECT 101043235 and supported by project

KLIMA-MEMES funded by the Bavarian Research Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities. The authors are responsible for the content of this publication.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Hadi Amiri, Timothy Miller, and Guergana Savova. 2018. Spotting Spurious Data with Neural Networks. 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 2006–2016, New Orleans, Louisiana. Association for Computational Linguistics.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. 2019. Unsupervised Label Noise Modeling and Loss Correction. In *Proceedings of the 36th International Conference on Machine Learning*, pages 312–321. PMLR.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *ACM Web Science 2013*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

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

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Markus Dickinson and W. Detmar Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. 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.

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. 2019. O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3325–3333.

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.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. 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 517–527, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Will Cukierski, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Raje, Max Bartolo, Evan Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Addison Howard, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, D. Sculley, Tzu-Sheng Kuo, Jonas W Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Y Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. 2023. DataPerf: Benchmarks for Data-Centric AI Development. In *Advances in Neural Information Processing Systems*, volume 36, pages 5320–5347. Curran Associates, Inc.

Mohammad Motamedi, Nikolay Sakharnykh, and Tim Kaldewey. 2021. A data-centric approach for training deep neural networks with less data. *CoRR*, abs/2110.03613.

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

Rodrigo Frassetto Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying Incorrect Labels in the CoNLL-2003 Corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226, 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.

Susanna Rücker and Alan Akbik. 2023. CleanCoNLL: A nearly noise-free named entity recognition dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8628–8645, Singapore. 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. *CoRR*, abs/1910.01108.

Lars Schmarje, Vasco Grossmann, Tim Michels, Jakob Nazarenus, Monty Santarossa, Claudius Zelenka, and Reinhard Koch. 2024. Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality. In *Pattern Recognition*, pages 459–475, Cham. Springer Nature Switzerland.

Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. 2023. Metadata archaeology: Unearthing data subsets by leveraging training dynamics. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? Evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on 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.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Vijay Vasudevan, Benjamin Caine, Raphael Gontijo Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. 2022. When does dough become a bagel? Analyzing the remaining mistakes on ImageNet. In *Advances in Neural Information Processing Systems*, volume 35, pages 6720–6734. Curran Associates, Inc.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training Named Entity Tagger from Imperfect Annotations. 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 5154–5163, Hong Kong, China. Association for Computational Linguistics.

Leon Weber and Barbara Plank. 2023. ActiveAED: A human in the loop improves annotation error detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8834–8845, Toronto, Canada. 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.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Jeff Huang, Chuyue Sun, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2023. Efficiently programming large language models using sglang. *arXiv preprint arXiv:2312.07104*.

A Data Statistics

A.1 Pair-wise inter-annotator agreements (Cohen's kappa) with MASI-distance for non-validated, self-validated, and peer-validated versions

versions \ *annotators*	1-vs-2	1-vs-3	1-vs-4	2-vs-3	2-vs-4	3-vs-4
before validation	0.40	0.42	0.37	0.36	0.31	0.34
self-validated	0.60	0.53	0.61	0.44	0.47	0.47
peer-validated	0.66	0.72	0.67	0.64	0.68	0.68

Table 4: Pair-wise inter-annotator agreements (Cohen's kappa) with MASI-distance for non-validated, self-validated, and peer-validated versions.

A.2 Frequency of NLI label on non-validation, self-validated, and peer-validated explanation-label pairs

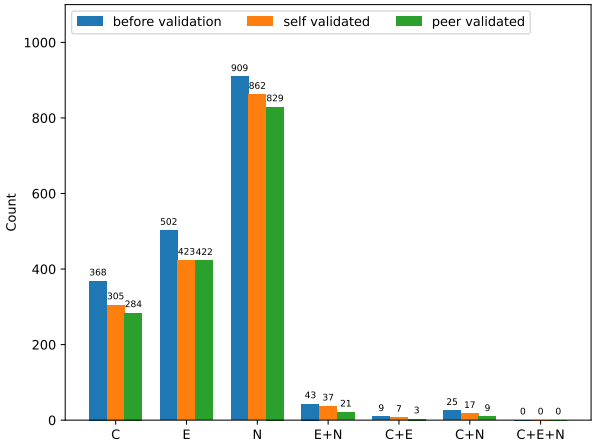


Figure 5: Frequency of NLI label sets on non-, self- and peer-validated label-explanation pairs.

B GPT Prompt

id: 72870c

System: You are an expert linguistic annotator.
User: We have collected annotations for a NLI instance together with reasons for the labels.
Your task is to judge whether the reasons make sense for the label.
Provide the probability (0.0 - 1.0) that the reason makes sense for the label.
Give ONLY the reason and the probability, no other words or explanation.
For example:
Reason: <The verbatim copy of the reason>
Probability: <the probability between 0.0 and 1.0 that the reason makes sense for the label, without any extra commentary whatsoever; just the probability!>.
Context: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.
Statement: Newspaper preprints can cost as much as $5.
Reason for label entailment: 5 dollars for a piece of newspaper
Reason for label neutral: The context only mentions how low the price may be, not how high it may be.
Reason for label neutral: The maximum cost of newspaper preprints is not given in the context.
Reason for label contradiction: The context says 5 or 6 cents, not $5.
User: *Reason*: 5 dollars for a piece of newspaper
Probability:
Assistant: 0.0
User: *Reason*: The context only mentions how low the price may be, not how high it may be.
Probability:
Assistant: 0.9
User: *Reason*: The maximum cost of newspaper preprints is not given in the context.
Probability:
Assistant: 0.8
User: *Reason*: The context says 5 or 6 cents, not $5.
Probability:
Assistant: 0.9

Figure 6: GPT Prompt for predicting likelihood probability of label-explanation pairs.

C More VariErr Examples

Premise: Students of human misery can savor its underlying sadness and futility.
Hypothesis: Students of human misery will be delighted to see how sad it truly is.
Label-explanation pairs: before validation: {E:1,N:2,C:1} Self-validated : {E:1,N:1} Peer-validated: {N:1}
Labels: [E, N] *Error:* [C]

		Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	A	Explanation	1	2	3	4
E	2	"can savor" implies "will be delighted".	✓	✓	×	×
N	1	It is not clear from the context if the students will be delighted.	×	×	✓	✓
	3	Students of human misery can "savored" that sadness, so maybe they are delighted to see that, maybe they are tortured by the disasters.	×	×	✓	✓
C	4	Savor means to understand. Not to enjoy.	×	×	?	×

(a) *id:* 116176c

Premise: The tree-lined avenue extends less than three blocks to the sea.
Hypothesis: The sea isn't even three blocks away.
Label-explanation pairs: before validation: {"E":4,"N":1,"C":1} Self-validated: {"E":3,"N":1} Peer-validated: {"E":4,"N":1}
Labels: [E, N] *Error:* [C]

		Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	A	Explanation	1	2	3	4
E	1	Both premise and hypothesis talk about less than three blocks.	✓	✓	✓	×
	2	If the avenue reaches the sea after less than three blocks, it cannot be further away.	✓	✓	✓	×
	3	The avenue is less than three blocks to the sea.	✓	✓	✓	×
	4	If the hypothesis means that the sea is less than three blocks away.	?	✓	✓	×
N	3	It is not given where is the location of the narrator.	✓	×	✓	✓
C	4	If the hypothesis means that the sea is more than three blocks away.	?	×	?	×

(b) *id:* 80630e

Premise: As he stepped across the threshold, Tommy brought the picture down with terrific force on his head.
Hypothesis: Tommy hurt his head bringing the picture down.
Label-explanation pairs: before validation: {"E":3,"N":1,"C":1} Self-validated: {"E":3,"N":1} Peer-validated: {"E":3,"N":1}
Labels: [E, N] *Error:* [C]

		Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	A	Explanation	1	2	3	4
E	1	the picture hit Tommy in the head	✓	✓	✓	×
	2	a picture hit Tommy's head with terrific force	✓	✓	✓	×
	3	Tommy hurt his head with the picture	✓	✓	✓	×
N	3	ambiguous if Tommy hurt himself or another guy	✓	✓	✓	×
C	4	Tommy is not hurt but rather bad strong emotion	×	×	✓	×

(c) *id:* 77893n

Table 5: Additional sample annotations from VARIERR NLI corpus. L: Label, A: Annotator; E: Entailment, N: Neutral, C: Contradiction; magenta: self-judgments, black: peer-judgments, Err: label error.