Understanding and Predicting Human Label Variation in Natural Language Inference through Explanations

Nan-Jiang Jiang¹ Chenhao Tan² Marie-Catherine de Marneffe¹³

¹ Department of Linguistics, The Ohio State University, USA

² University of Chicago, USA

³ FNRS, UCLouvain, Belgium

jiang.1879@osu.edu chenhao@uchicago.edu demarneffe.1@osu.edu

Abstract

Human label variation (Plank, 2022), or annotation disagreement, exists in many natural language processing (NLP) tasks. To be robust and trusted, NLP models need to identify such variation and be able to explain it. To this end, we created the first ecologically valid explanation dataset with diverse reasoning, LIVENLI. LIVENLI contains annotators' highlights and free-text explanations for the label(s) of their choice for 122 English Natural Language Inference items, each with at least 10 annotations. We used its explanations for chain-of-thought prompting, and found there is still room for improvement in GPT-3's ability to predict label distribution with in-context learning.

1 Introduction

Until recently, practices for operationalizing annotations in natural language processing (NLP) datasets assumed one single label per item. However, human label variation (Plank, 2022) has been found in a wide range of NLP tasks, including part-of-speech tagging, coreference resolution, and natural language inference (NLI) (Plank et al., 2014; Poesio et al., 2018; Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Uma et al., 2021; Jiang and de Marneffe, 2022, i.a.). Aroyo and Welty (2015), among others, argued that such variation in annotations should be considered signal, not noise. Specifically, the NLI task — identifying whether the hypothesis is true (Entailment), false (Contradiction), or neither (Neutral) given a premise — has embraced label variation and set out to predict it (Zhang et al., 2021; Zhou et al., 2022). However, the question of where label variation in NLI stems from remains open.

As an initial effort to tackle this question, Jiang and de Marneffe (2022) introduced a taxonomy of linguistic phenomena that can lead to different interpretations of English NLI items, and thus potentially to different NLI labels. For instance, in Ex.1 (Table 1), does one infer *a large following* in the hypothesis from *most* in the premise? This lexical indeterminacy was hypothesized by Jiang and de Marneffe (2022) to have contributed to the variation that is reflected in the label distribution.

Jiang and de Marneffe (2022)'s taxonomy was built post-hoc, and hence detached from the annotators who provided the NLI labels. It is thus important to understand the reasons from the perspective of annotators and validate their taxonomy. Furthermore, given that human label variation is widespread and systematic, NLP models that can predict and explain label variation would be useful for improving annotation and downstream utility.

We introduce LIVENLI, Label Variation and Explanation in NLI. LIVENLI contains 122 reannotated MNLI (Williams et al., 2018) items, each with at least 10 annotations, including highlights and free-text explanations for the labels chosen by the annotators. Compared to previous explanation datasets, LIVENLI offers more diverse and ecologically valid explanations as each annotator explained the label they chose.

LIVENLI provides direct evidence for how label variation among annotators arises, allowing us to compare annotators' explanations against Jiang and de Marneffe (2022)'s taxonomy. The free-text explanations in LIVENLI showed that the reasons for label variation are largely similar to the taxonomy, but other reasons also emerge. Further, we found within-label variation: even when annotators chose the same label, they may have different reasons to do so. Our highlights analysis also confirmed Tan (2022)'s observation that highlights by themselves do not provide effective explanations.

Finally, we explore the promise of using NLP models to predict human label variation, and LIVENLI allows us to introduce explanations as additional supervision. Inspired by chain-of-

Ex.1 P: Most pundits side with bushy-headed George Stephanopoulos (This Week), arguing that only air strikes would be politically palatable. H: Mr. Stephanopoulos has a very large pundit following due to his stance on air strikes only being politically palatable.

LIVENLI [E,N,C]: [0.4, 0.3, 0.3]

QUD: Does Stephanopoulos have a very large pundit following?

Resp.1.1: Entailment – This statement is most likely to be true because in the context is stated that "Most pundits" would side with Mr. Stephanopoulos. Most pundits could also mean a very large pundit following.

Resp.1.2: Neutral – You cannot infer that the overall number of pundits following the individual is large just because the majority of pundits follow the individual. He could just have 2 out of 3 total pundits following him, for instance. Furthermore, they may be following him for reasons outside his stance on air strikes.

QUD: Do pundits follow Stephanopoulos due to his stance on air strikes?

Resp.1.3: Neutral – George Stephanopoulos may have a follow from pundits, but it might not be due to his support of drones.

Resp.1.4: Contradiction – He might have a large pundit following, but that would have to be for something before the current issue of air strikes since one event wouldn't get people a large following overnight.

Ex.2 **P:** They encourage the view that there's nothing—from Iraqi germ weapons programs to Serbian atrocities—that a few invisible planes can't fix. **H:** A few invisible planes are all it takes to fix certain issues. LIVENLI [E,N,C]: [0.5, 0.35, 0.15] QUD: Can a few invisible planes fix certain issues?

Resp.2.1: Neutral | Contradiction – The context is unclear because "They are encouraging the view" does not make the following part of the context true or false.

QUD: Are a few invisible planes all it takes to fix certain issues?

Resp.2.2: Neutral – Invisible planes may solve some problems but I'm not sure they are all it takes to fix certain issues. Could be true or false.

Ex.3 P: for a change i i got i get sick of winter just looking everything so dead i hate that **H:** I'm so sick of summer.

LIVENLI [E,N,C]: [0, 0.35, 0.65]

Resp.3.1: Contradiction – The context is stating how one is sick of winter, not summer, as the statement describes.

Resp.3.2: Contradiction – The speaker hates winter because the foliage is dead, therefore he likely loves summer when everything is alive.

Resp.3.3: Neutral – The context mentions being sick of winter while the statement mentions being sick of summer. These could both be true because the same person may still complain of summer's heat.

Ex.4 **P:** The original wax models of the river gods are on display in the Civic Museum. **H:** Thousands of people come to see the wax models. original MNLI: [0,1,0] LIVENLI [E,N,C]: [.23, .73, .04]

Resp.4.1: Neutral – The context refers to the wax model displays in the museum. The context makes no mention of the number of visitors mentioned in the statement.

Resp.4.2: Entailment – Museums are generally places where many people come, so if the original wax models are there it is likely thousands of people will come to see them.

Resp.4.3: Entailment | Contradiction – It's unlikely a museum could stay open for very long without thousands of visitors, so it's likely true that thousands of people come to see these wax mdoels. Unless, of course, it's a big museum with many attractions more interesting than the models, in which case the statement is likely to be false.

Table 1: Examples in LIVENLI. P: Premise. H: Hypothesis. [E,N,C]: the probability distributions over the labels (E)ntailment, (N)eutral, and (C)ontradiction, aggregated from the multilabel annotations in LIVENLI. The explanations in the Responses (Resp.) help contextualize the label distributions.

thought prompting (Wei et al., 2022), we show that incorporating explanations in the prompt slightly improves the ability of GPT-3 (Brown et al., 2020) to predict NLI label distributions over prompting without explanations. Yet, there is still room for improvement in both predicting distributions and generating explanations.¹

2 Related Work

2.1 Label Variation in NLI

Previous work has found systematic label variation in NLI datasets (de Marneffe et al., 2012; Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Uma et al., 2021). Jiang and de Marneffe (2022)

proposed a taxonomy of 10 linguistic categories that can lead to label variation. Ex.1 illustrated the "Lexical" category. We refer the reader to Jiang and de Marneffe (2022) for the full definition of all 10 taxonomy categories.

In this work, we validate and improve the taxonomy based on the explanations annotators provide when choosing their label(s). The taxonomy categories do not specify what the different readings are for the different labels, e.g., in the "Lexical" category, which words carry uncertain meaning. Another frequent taxonomy category in Jiang and de Marneffe (2022) is "Probabilistic Enrichment": some annotators make different probabilistic inferences from the premise that are likely but not definitely true. In Ex.3 (Table 1), Responses 3.2 and

¹The data and code will be publicly available on GitHub.

3.3 make different inferences about whether the speaker likes summer given that they dislike winter: 3.2 considers that the speaker likes it because of the live foliage, while 3.3 hypothesizes that the speaker might dislike summer because of the heat. The category "Probabilistic Enrichment" by itself does not tell us what the probabilistically enriched content is. The explanations, on the other hand, do convey this information, and would be more useful for non-experts to understand label variation.

Predicting NLI label distribution Pavlick and Kwiatkowski (2019) argued that models should preserve information about label variation by predicting label distributions to propagate such information to downstream tasks. Zhang et al. (2021) and Zhou et al. (2022) introduced BERT-based fine-tuned models for predicting NLI label distribution, using the ChaosNLI dataset (Nie et al., 2020) which contains 100 annotations/item. Here, we explore predicting label distributions with in-context learning with large language models, namely GPT-3 (Brown et al., 2020), using explanations as additional supervision.

2.2 Explanation Datasets

With the rise of interest in interpretability in NLP, many datasets where labels are given explanations have been introduced (see Wiegreffe and Marasovic (2021) for a survey). The dataset most similar to LIVENLI is e-SNLI (Camburu et al., 2018): it also targets the NLI task and contains free-text explanations and highlights for all items in SNLI (Bowman et al., 2015). e-SNLI has been extensively used for building models that produce explanations (Kumar and Talukdar, 2020; Narang et al., 2020; Zhao and Vydiswaran, 2021; Wiegreffe et al., 2022; Yordanov et al., 2021, i.a.) and studying whether models can learn from explanations to perform the classification task (Wiegreffe et al., 2021; Hase and Bansal, 2022; Ye and Durrett, 2022, i.a.). Our work extends e-SNLI on several dimensions.

Ecological validity An explanation is ecologically valid if the same annotator provides both the label to explain and the explanation. Many explanation datasets are not ecologically valid: instead of soliciting both the label and the explanation, only the explanations for "why this item has this label" were collected, and the label to explain is the "ground truth" label collected in previous work. However, given that label variation

is widespread in many NLP tasks, only explaining the ground truth is not capturing the full range of meaning and may not reflect the annotators' decision making process. This also poses a problem for evaluating explanations, as Wiegreffe et al. (2022) found that humans have biases against explanations when they disagree with the label. In this work, we ask annotators to label NLI items and explain their chosen label, ensuring the ecological validity of the explanations.

Diversity of the explanations We aim to collect explanations that are diverse on multiple aspects. First, as discussed earlier, most existing explanations seek to explain the "ground truth" label without taking into account human label variation. By focusing on explaining items that are known to exhibit variation, explanations in LIVENLI reflect how the same item can have different labels.

Second, most explanation datasets for NLP tasks with free-text explanation contain only one explanation per item. The few large scale datasets with more than one explanation have up to 5 explanations/item (Camburu et al., 2018; Sap et al., 2020; Zhang et al., 2020; Do et al., 2020). For LIVENLI, we collected 10 explanations per item, often explaining more than one label. This provides us with rich information about reasons behind each label. As we will see, people can indeed arrive at the same label for different reasons.

Lastly, LIVENLI is based on MNLI (Williams et al., 2018), whereas e-SNLI is based on SNLI. As pointed out by Williams et al. (2018), SNLI describes visual scenes only, making the hypotheses short and simple, and non-visual linguistic phenomena, e.g. temporal reasoning and beliefs, rare. Therefore, the e-SNLI explanations also tend to be simple and templated (Camburu et al., 2020; Tan, 2022), and are unlikely to exhibit diverse types of reasoning. We chose MNLI to ensure that the items involve different types of reasoning, which in turn contributes to having diverse explanations.

Learning from explanations Previous work did not find much success incorporating explanations to improve classifiers using datasets such as e-SNLI with fine-tuning settings (Camburu et al., 2018; Wiegreffe et al., 2021; Hase and Bansal, 2022). Chain-of-thought prompting, a.k.a. Explain-then-Predict (Wei et al., 2022; Ye and Durrett, 2022), on the other hand, improves large language model's ability to perform some reason-

ing tasks when prompting the model with explanations for the answer before generating it. We explore using LIVENLI's explanations to improve GPT-3's ability to predict label variation.

3 Data Collection

We re-annotated 122 items from the MNLI dev set: 110 are analyzed in Jiang and de Marneffe (2022) – 50 of them are also in the ChaosNLI-re-annotation (Nie et al., 2020), with 100 annotations/item. The other 12 items (4 for each NLI label) have unanimous agreement in the original MNLI 5 annotations. We intended to use those for quality control, to filter annotators who did not give the same label as the original one and might therefore not be paying attention. However, as discussed below, those items also exhibit systematic label variation with legitimate explanations.

Procedure Figure 1 shows a screenshot of the annotation interface on the Surge AI crowdsourcing annotation platform.² Annotators read the premise (context) and hypothesis (statement), and were asked whether the statement is most likely to be true/most likely to be false/either true or false (corresponding to Entailment/Contradiction/Neutral, respectively). They can choose multiple labels if they feel uncertain. They were then asked to provide a free-text explanation for all the label(s) they chose. They were instructed to give explanations that provide new information and refer to specific parts of the sentences, and avoid to simply repeating the sentences. These instructions were inspired by previous work that identified features of high and low quality explanations (Camburu et al., 2020; Wiegreffe et al., 2022; Tan, 2022). Annotators were further instructed to highlight words from the premise and hypothesis that are most important for their explanations. Surge AI guarantees that their annotators are native English speakers, as well as their reading and writing abilities. Annotators were paid \$0.6 per item (hourly wage: \$12 to \$18). Each item received at least 10 annotations. 48 annotators participated in total. On average, each annotator labeled 29.3 items. Our data collection was performed with IRB approval.

4 Label Analysis

We analyzed the multilabel annotations to quantify the amount of variation and the extent to which they align with the original MNLI annotations.

Descriptive statistics In 83% of the label responses, only one label was chosen, with 53% of all responses being Entailment or Contradiction. The majority of the annotators thus express clear judgments. Much of the label variation arises from variation across individuals as opposed to each individual having uncertain or gradient judgments.

To aggregate the labeling responses for each item, we first normalize each multilabel response into a distribution, so that if an annotator chose multiple labels, each label will be weighted less than if they only chose a single label. We then average the individual distributions to obtain the item's label distribution. In Figure 2, each set of stacked bars shows the distributions for one item. The transparency of each bar indicates the mean non-zero probabilities for that label. Figure 2 shows that there are fine-grained judgments that we need to predict in the form of label distributions. On the other hand, there can be similar distributions (shown by the lengths of the bars) coming from different compositions of multilabels (shown by the transparency of the bars).

Agreement items We intended to use 12 items, originally with unanimous agreement in MNLI, for quality control. For each of these 12 items, the labels with the highest probability in the aggregated distribution from our annotations are the same as the unanimous one in the original MNLI annotations, suggesting that our annotators' judgments are reliable. However, as the right panel of Figure 2 shows, there is a wide range of variation for those items, but annotators provided legitimate explanations for the labels that differ from the original one (see Ex.4 in Table 1).

Takeaways We found that there is systematic label variation in the LIVENLI annotations, including for items that were unanimously agreed on in the original MNLI annotations. This suggests that label variation is widespread, and we need more annotations to capture it. Baan et al. (2022) pointed out that considering one distribution to govern all judgments from the population is a simplification. Our findings confirm that the distributions collected from a pool of annotators

²surgehq.ai

Read the following context and statement: Context: Could you please speak to this issue, with regard to the social ramifications of gum chewing in public? Statement: You don't have an opinion on gum chewing in public, I see. Choose one or more from the following: If you feel uncertain and you feel that multiple options apply, choose them all instead, even though it might feel contradictory. Assuming the context is true, the statement: is most likely to be true can be either true or false is most likely to be false Explain, in a few sentences, why you chose your answer. If you chose more than one option, elaborate in which circumstances each option is possible. Explain all the options you chose. Your explanation should include new information and refer to specific parts of the sentences. It should NOT simply repeat the sentences Avoid "The context and statement means the same/opposite thing". Specify which part of the context and statement means the same/opposite thing Avoid "Just because X doesn't mean Y". Say under what circumstances X does not mean Y, or say that X can mean Y or Z. Avoid "The statement is ambiguous/it's not clear what it means". Elaborate what the possible meanings are and why it is ambiguous.

Highlight the words in the Context and Statement that are relevant to your explanations.

Your explanations should refer to specific words/parts of the sentences. Highlight those words and phrases that your explanations mentioned. Only highlight the words that are most important for the explanations.

Figure 1: Data collection interface.

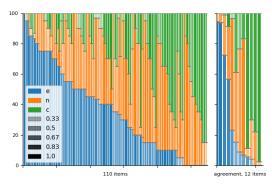


Figure 2: Average distributions from the normalized multilabel responses for each item. For each item and each label, we take the multilabel responses that include at least that label and compute the average probabilities for that label (bars' transparency). Label variation is widespread, even for the items with unanimous agreement in MNLI.

(a sample of the population) may shift with a different pool.³

5 Free-text Explanations Analysis

Jiang and de Marneffe (2022) proposed a taxonomy of reasons for label variation in NLI. Here we qualify the diversity of the LIVENLI explanations, and analyze the explanations to see what are annotators' reasons for label variation – and in particular if they are similar to what Jiang and de Marneffe (2022) hypothesized.

Minimum word count: 10 Words: 0

5.1 Quantitative Analysis

For each item i with each (multi)label l, Figure 3 shows the entropy of unigram distributions calculated using i's explanations for l, and all the subsets of those explanations (sizes on the x-axis), with error bars indicating the variability across items i. We also perform the same calculation for the explanations from the e-SNLI dev and test sets. Both datasets have the same ordering in the entropy of explanations for single labels: Neutral > Contradiction > Entailment. This is expected: to explain Entailment, annotators reiterate the premise/hypothesis and tend to use the same words used in the sentences (as in Response 1.1), while to explain Neutral and Contradiction, an-

³Comparing with the label distributions in ChaosNLI reinforces this statement. We omit the analysis with ChaosNLI here for space reasons.

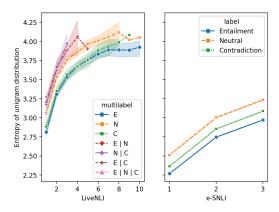


Figure 3: Entropy of distributions of unigrams in the explanations of each item with each label (error bars indicating the variability across items.)

notators need to mention information not already given and are thus likely to use different words (as in Responses 1.2 and 1.3).

Compared with e-SNLI, our explanations have higher unigram distributions entropy, even for the same number of explanations (3/item), showing that our explanations are more lexically diverse. They are also longer, with mean token length 28.7 vs. 13.3 for e-SNLI. Focusing on the left panel, we see that entropy increases with the number of explanations: it is not surprising that more explanations increases lexical diversity. But importantly, the entropy plateaus after 5 explanations for all labels, which could be the sweet spot for balancing between diversity and annotation budget.

5.2 Verifying the Taxonomy

The explanations in LIVENLI allow us to investigate the reasons for which annotators chose different labels. We investigate how often Jiang and de Marneffe (2022)'s categories for label variation appear in the explanations. For each pair of premise/hypothesis, one author examined its explanations and assigned the pair one or more categories of variation exhibited in its explanations.⁴

Table 2 shows the frequency of the taxonomy re-annotations. We can see that all the taxonomy categories from Jiang and de Marneffe (2022) are used. Across the board, label variation arises in similar reasons as what Jiang and de Marneffe (2022) hypothesized. Notably, the two most frequent categories are Probabilistic and Lexical.

Meanwhile, we found that many instances of label variation stem from annotators judging the

| Combination of Taxonomy Categories | Frequency | |
|------------------------------------|-----------|--|
| Probabilistic | 28 | |
| Lexical | 21 | |
| QUD | 15 | |
| Coreference | 6 | |
| Lexical Probabilistic | 5 | |
| Coreference Probabilistic | 5 | |
| Imperfection Lexical | 5 | |
| Implicature | 5 | |
| Imperfection Probabilistic | 4 | |
| Probabilistic QUD | 4 | |
| No variation | 3 | |
| Coreference Lexical | 3 | |
| Lexical QUD | 3 | |
| Temporal | 2 | |

Appear once:

Implicature | Interrogative; Implicature | Lexical; Interrogative; Implicature | QUD; Imperfection | QUD; Interrogative | Lexical; Coreference | Temporal; Coreference | Imperfection; Coreference | QUD; Implicature | Interrogative; Implicature | Lexical; Interrogative; Implicature | Lexical | Probabilistic | QUD; Probabilistic | Temporal; Presupposition | QUD; QUD | Temporal

Table 2: Frequency of the taxonomy categories in LIVENLI. We italicize the new category that emerge from LIVENLI. Label variation in annotators arises for similar reasons as what Jiang and de Marneffe (2022) hypothesized.

truth of different at-issue content, which answers different Questions Under Discussion (QUDs) (Roberts, 2012). For example, in Ex.1, Responses 1.1 and (the first half of) 1.2 take the main point of the hypothesis to be Stephanopoulos having a very large pundit following, but have different judgments on whether this main point is true. On the other hand, 1.3 and 1.4 focus on the reason for which pundits follow Stephanopoulos, and agree with 1.1 that he has a large following. We thus added a "QUD" category to the taxonomy (generalizing two of Jiang and de Marneffe (2022)'s categories, "accommodating minimally added content" and "high overlap", which involve annotators having different readings and ignoring certain parts of the items). The QUD category occurred in 28 items (out of 122) and is the third most frequently used category. QUD is thus an important aspect of language understanding that people pick up on. Incorporating QUD into NLP task designs and modeling (De Kuthy et al., 2018; Narayan et al., 2022) is an interesting avenue of research.

5.3 Within-Label Variation

Even though the taxonomy categories were meant to be reasons for variation in labels, LIVENLI's

⁴For some items, annotators unanimously agree on one label and on their reasons for the label. Those are assigned "No variation" in Table 2.

explanations show that there is also *within-label variation*: annotators can vary in their understanding of the text and have different reasons for the same label. 16 (out of 122) items exhibit such within-label variation, which the taxonomy can capture. We discuss here two categories in which within-label variation occurs.

Different QUDs In Ex.2 (Table 1), some annotators agree on the Neutral label but their explanations show that they take the main point of the sentence to be different: Response 2.1 focuses on whether the view being encouraged is true, while 2.2 takes the view to be true while questioning whether the planes are all it takes.

Different coreference assumptions In Ex.3, the explanations indicate that the annotators differ in whether to take the premise and hypothesis to be referring to the same entity/event. Response 3.1 assumes that the premise and hypothesis refer to the same season: thus since the premise talks about winter and not summer, the hypothesis *I'm so sick of summer* is false. 3.2, on the other hand, does not assume that the premise and the hypothesis refer to the same season, and infers through probabilistic enrichment that the speaker likes summer, making the hypothesis false. Both with and without the coreference assumption, annotators label the item as Contradiction.

To summarize, explanations reveal the kinds of variation that labels themselves do not capture. LIVENLI will be a useful resource to study variation in human understanding of texts in general, not limited to variation in labels.

6 Highlights Analysis

Highlights have been criticized in previous work as an inadequate form of explanation (Tan, 2022) because they only provide evidence for the label without conveying the mechanisms for how the evidence leads to the label. Here, we evaluate the quality of the highlights in LIVENLI, and analyze what information they convey.

Relationship between highlights and labels One might hypothesize that annotators would agree on the highlights when they agree on the label. To measure agreement on highlights, for each item, we computed the overall Krippendorff's α on whether annotators highlight each word using all the highlights for the item, as well as 7 label-specific α 's using the subset of highlights that

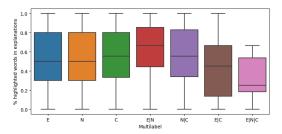


Figure 4: Proportion of highlighted words that are also mentioned in the explanation.

agreed on the (multi)label.⁵ The mean overall α across all items is 0.22, while the label-specific α 's range from 0.25 to 0.31, suggesting low agreement. In Ex.5 and Ex.6 (Table 3), annotators unanimously agreed on the label while giving different highlights. Sullivan Jr. et al. (2022) similarly found low agreement on highlights for sentiment analysis (α = 0.3), even though the agreement on the sentiment label was high (α close to 1).

On cases where annotators do agree on the highlights, they can disagree on the label. In fact, the same highlights can be given for different labels. In Ex.7, three annotators provided the exact same highlights (only the words *simile* and *metaphor*), while providing three different labels. They all consider the words *simile* and *metaphor* to be important but have different understanding of the relationship between similes and metaphors. This suggests that, for variation in NLI, the label depends not only on which words are important, but crucially on the relationship between the words.

How grounded in the explanations are the highlights? Figure 4 shows the boxplot of highlights

lights? Figure 4 shows the boxplot of highlights and explanation overlap (percentages of words highlighted that appear in the corresponding explanation) organized by the (multi)label they explain. On average, the explanations mention 57% of the highlighted words. When annotators chose all the labels (EINIC), the percentage of highlighted words is much lower: the free-text explanations mention that the annotators find the sentences hard to understand and that they cannot make an informed judgment. This further suggests that the highlights are noisy. Many highlights include large spans of words covering entire clauses, while only a few words in the clause are discussed in the explanation. Sullivan Jr. et al. (2022) found that removing drag affordances from the user in-

 $^{^5}$ 155 responses were not included to calculate the label-specific α s because there is no other responses for the respective item choosing the same multilabel.

Agree on label, not on highlights Ex.5 **P**: There are other reasons that wrecks cause fan excitement—e.g. **H**: A cause for fan excitement can be wrecks. Ex.6 **P**: He leaned over Tommy, his face purple with excitement. **H**: He hovered over Tommy, with a deep color in his face from the thrill. Ex.5 **P**: There are other reasons that wrecks cause fan excitement—e.g. **H**: A cause for fan [E,N,C]: [1, 0, 0] $\alpha = 0.08$

Agree on highlights, not on labels

Ex.7 **P**: What a brilliantly innocuous metaphor, devised by a master manipulator to obscure his manipulations. **H**: The simile was created by the manipulator. [E,N,C]: $[0.3, 0.4, 0.3] \alpha = 0.43$

Entailment – As all metaphors are a subset of similes, the statement must be true.

Neutral – A metaphor is not the same as a simile, which the manipulator definitely created, but they could have created a simile too.

Contradiction – The context describes a metaphor, while the statement mentions a simile, which is a different type of speech.

Table 3: Examples where agreement on either the label or the highlights does not imply agreement on the other. Color transparency indicates the percentage of annotators who highlighted the word. The agreement on highlights do not correlate with agreement on labels.

terface (forcing annotators to click each word individually) increases inter-annotator agreement. We used a top-of-the-shelf web interface provided by Surge AI that allows drag affordances, which may have contributed to the noise we found.

What leads to high agreement on highlights?

The mean highlights and explanation overlap for each item is positively correlated with the highlight overall α for each item (Pearson's r=0.38). The items with high overlap and high α tend to have hypotheses with parallel structures to the premise while differing minimally in certain words: the pair of words that differ are then highlighted and discussed in the explanations. For example, *leaned-hovered* in Ex.6 and *metaphorsimile* in Ex.7 are consistently highlighted.

Takeaways Our analysis reiterates previous criticism that highlights do not effectively explain labels. In particular, the same highlight can be used to explain different labels. For complex tasks such as label variation in NLI, the free-text explanations are crucial for explaining what the relationships are between the highlighted words and between the highlights and the labels.

7 Predict Label Variation by Learning from Explanations

Given that there is systematic label variation in data, models need to be able to recognize such variation to obtain human-like understanding. Explanations have been found to improve large language models' performance on various tasks (Wei

et al., 2022). Therefore, we investigate to what extent LIVENLI's explanations can be used to help large language models identify label variation.

Setup We use the latest variant of GPT-3 (Brown et al., 2020): text-davinci-003 with the in-context learning paradigm. We prompt the model with training items, including the premise/hypothesis, the label distribution in prose, and the explanations for each label, followed by a test item, consisting of the premise/hypothesis only, without any label distribution or explanations. Table 4 shows the prompt for one training item.

We experiment with two orderings of the prediction and explanation in the prompt: Predict-then-Explain vs. Explain-then-Predict (a.k.a. chain-of-thought prompting). Ye and Durrett (2022) found no clear winner between the two approaches for reasoning tasks like NLI. Each prompt includes 16 training items, which is the highest number of items that fit within the length limit when the prompt includes explanations. We use the longest explanation for each label for training to increase lexical variability.

To investigate the effectiveness of explanations, we also prompt the model without any explanations (Predict-only) with the same 16 training items. However, since the models prompted with explanations receive strictly more information, we also experiment with providing additional 16 training items (Predict-only-extra-train) to compensate for the loss of information.

We first developed the prompts on 20 items that

Context: Although it is a significant part of the poverty population, Asians historically have not been able to participate in the services and programs available to the poor, he said. **Statement:** Asians are usually not poor.

Prediction | Given the context, the probability of the statement being true is 0.0, the probability of the statement being false is 0.85, the probability of the statement being either true or false is 0.15.

Explanations It can be false because the context notes that Asians are typically a significant part of the poverty population whereas the statement notes that Asians are not usually poor. These two statements seem to contradict one another. Thus, the statement is most likely false.

It can be either because the context says Asians make a significant part of the poverty population, which would imply that the statement itself is false if the term significant referred to the overall number. However, since there are more Asians than any other demographic, it could just as likely be that they are usually not poor but just make up a significant portion due to them making up a significant portion of the overall population.

Table 4: Training example prompt for the Predict-then-Explain task. The texts Prediction and Explanations are not part of the prompt. The Explain-then-Predict prompt flips the order of Prediction and Explanations. The Predict-only prompt does not include Explanations.

are not used for training or testing. For the remaining 102 items, we created three train/test splits of sizes 32/70. We report the mean and standard deviation over the three test sets.

Metrics To measure the distance between the predicted and the ground truth label distributions, we report KL-divergence $([0,\infty))$, Jenson-Shannon distance (JSD) and Total Variation Distance (TVD) (both bounded within [0,1]). The lower the metrics, the better the predictions. We include a baseline of always predicting the uniform distribution with 1/3 probability for each label (Uniform in Table 5). As a reference, to see how much variation needs to be captured, we also predict the ground truth majority vote label of each item with probability 1 (Majority Vote in Table 5).

Results Table 5 shows the metrics for the different setups, as well as the entropy of the predicted distributions.⁶ Overall, the KL and JSD results are comparable to Nie et al. (2020), who finetuned BERT-based (Devlin et al., 2019) classi-

| | JSD | TVD | KL | Entropy |
|--------------------------|-----------------|-----------------|-----------------|------------------------|
| Explain-then-Predict | $0.272_{0.015}$ | $0.598_{0.022}$ | $0.682_{0.432}$ | 0.899 _{0.059} |
| Predict-then-Explain | $0.278_{0.034}$ | $0.629_{0.096}$ | $0.982_{0.322}$ | 0.6180.066 |
| Predict-only | $0.284_{0.016}$ | $0.626_{0.041}$ | $1.302_{0.518}$ | $0.579_{0.044}$ |
| Predict-only-extra-train | $0.268_{0.025}$ | $0.585_{0.051}$ | $1.147_{0.325}$ | $0.646_{0.022}$ |
| Uniform | 0.285 | 0.665 | 0.348 | 1.098 |
| Majority Vote | 0.363 | 0.697 | 4.066 | 0 |
| Nie et al. (2020) | 0.305 | | 0.665 | |
| Zhang et al. (2021) | 0.192 | | 0.180 | 0.868 |

Table 5: Metrics for each setup and mean entropy of the predicted distributions. Each value is the average of three random splits (standard deviation in subscript). Best results from Nie et al. (2020) and Zhang et al. (2021) on ChaosNLI given as reference – not directly comparable (with us and with each other) because different test sets are used.

fiers on single-labeled data only and evaluated the softmax distributions. Our results underperform Zhang et al. (2021), who also fine-tuned BERT-based models with both single-labeled and soft-labeled data, suggesting room for improvement.

Effect of explanations Explain-then-Predict slightly outperforms the setups with the same amount of training data. The difference in KL is larger but may not be meaningful (see below). Doubling the amount of training items improves some metrics: Predict-only-extra-train outperforms Explain-then-Predict on JSD and TVD.

The entropy of the prediction shows a larger difference between the models. In particular, Explain-then-Predict predicts high entropy distributions: predicting with lower probabilities for all labels, leading to better KL and TVD with the ground truth distributions. Predict-only tends to be over-confident: for 40% of the items, it incorrectly predicts some labels to have 0 probability. This suggests that when GPT-3 generates the explanations first, especially before the labels (as in Explain-then-Predict), it considers more labels to be possible, leading to more uniform distributions.

Difference in KL is not meaningful The goodness on the KL-divergence correlates with the entropy of the predicted distributions, with the Uniform baseline scoring best. Nie et al. (2020) reported a similar strong performance from the Uniform baseline. Instead of interpreting the pattern as suggesting that the uniform baseline is strong, we argue that the KL-divergence definition actually explains this pattern. KL-divergence of the ground truth y from the prediction \hat{y} KL $(y||\hat{y}) = H(y, \hat{y}) - H(y)$, where $H(y, \hat{y})$ is the cross-

⁶98% of the predicted distributions sum to 1. We normalize the distributions before computing the metrics. We also smooth the predictions by adding 1e-5 to all probabilities to prevent undefined KL on a predicted probability of 0.

1 Context: Although it is a significant part of the poverty population, Asians historically have not been able to participate in the services and programs available to the poor, he said. **Statement:** Asians are usually not poor.

It can be true because the context states that Asians are a significant part of the poverty population, which implies that they are not usually poor.

It can be either because the context does not specify whether or not Asians are usually poor or not, so it is impossible to know for sure.

It can be false because the context states that Asians are a significant part of the poverty population, which implies that they are not usually not poor.

Predicted [E,N,C]: [0.7, 0.1, 0.2] Gold: [0.0, 0.15, 0.85] KL = 1.29 JSD = 0.59 TVD = 0.7 Human explanations given in Table 4

2 Context: What a brilliantly innocuous metaphor, devised by a master manipulator to obscure his manipulations. **Statement:** The simile was created by the manipulator.

It can be true because the context stated that the master manipulator created an innocuous metaphor, which is another name for simile, indicating that the statement is true.

It can be either because the context indicates that the metaphor created was innocuous, which means it is not obviously deceptive or manipulative. Therefore, the statement could be true, meaning it was created by the manipulator, or false, meaning it was not created by the manipulator.

It can be false because the context does not say for sure that the metaphor was created by the manipulator, so it could be false. The context only assumes that, or suggests it.

Predicted [E,N,C]: [0.67, 0.18, 0.15] Gold: [0.3, 0.4, 0.3] KL = 0.29 JSD = 0.27 TVD = 0.37 Human explanations given in Table 3, example Ex.7

Table 6: Label prediction and generated explanations from Explain-then-Predict. The generated explanations can be implausible or inaccurately describe the input sentences.

entropy between y and \hat{y} , which is also the surprisal of \hat{y} , weighted by y. When a label x has nonzero ground truth probability y(x) > 0, the cross-entropy $-y(x)\log\hat{y}(x)$ is smaller when the surprisal of the prediction $-\log\hat{y}(x)$ is smaller, which is when $\hat{y}(x)$ gets larger. Therefore, a prediction that incorrectly predicts the label to have a 0 probability gets an exponentially larger KL than predicting the uniform probability. JSD, on the other hand, takes the average distribution of y and \hat{y} and therefore upweights the 0-probability predictions in \hat{y} and lessens the penalty. This suggests that the difference in KL-divergence may not be meaningful if models are performing worse than the uniform baseline, which is the case here.

Evaluation of the generated explanations We perform a qualitative analysis of the generated explanations, focusing on the output of Explain-

then-Predict on one test split containing 70 items and 218 explanations.⁷

We found that 32 items have at least one problematic explanation: 29 explanations are highly implausible and 7 explanations inaccurately describe what is stated in the premise/hypothesis. 16 of the problematic explanations are associated to the Contradiction label, and 13 to Entailment. The mean predicted probability of the labels with problematic explanations is 0.19, while the mean true probability of those labels is 0.10. This suggests that the model not only generates poor explanations but also predict those labels to be more likely than the ground truth.

Table 6 shows examples for both kinds of errors. Example 1 has an implausible explanation for the "true" (Entailment) label: being a significant part of the poverty population cannot imply that they are not usually poor. In Example 2, the explanation for "false" (Contradiction) inaccurately describes the context: the context does say that the metaphor was created (devised) by the manipulator. The generated explanations in both examples include template-like language, and do not describe much reasoning over the text, unlike the human explanations. We focus only on these two kinds of errors because they are straightforward to identify. Evaluating model-generated explanations is a challenging and open question in general. We leave the systematic evaluation of the generated explanations, perhaps using some of the criteria proposed by Wiegreffe et al. (2022), for future work.

8 Conclusion

We introduced LIVENLI, containing ecologically valid explanations for different NLI labels. We showed that the explanations are diverse in their lexical choices and in the reasons they convey. When prompting with GPT-3, there is still room for improvement in both predicting label distributions and generating explanations for different labels. We hope LIVENLI can be a useful resource for studying variations in text understanding and for building NLP models that capture variation.

⁷For 68 items, the model generated explanations for all three labels. For the remaining 2 items, it generated explanations for two labels. 5 explanations are for labels that were predicted to have 0 probability.

References

- Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

- Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-snli-ve-2.0: Corrected visual-textual entailment with natural language explanations. *CoRR*, abs/2004.03744.
- Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. 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.
- Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.
- 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.

- Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. *arXiv:2004.14546 [cs]*. ArXiv: 2004.14546.
- Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Dipanjan Das, and Mirella Lapata. 2022. Conditional generation with a question-answering blueprint.
- 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.
- 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. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.
- Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.
- Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5.

- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Jamar Sullivan Jr., Will Brackenbury, Andrew McNutt, Kevin Bryson, Kwam Byll, Yuxin Chen, Michael Littman, Chenhao Tan, and Blase Ur. 2022. Explaining why: How instructions and user interfaces impact annotator rationales when labeling text data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–531, Seattle, United States. Association for Computational Linguistics.
- Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. 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 Re*search, 72:1385–1470.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.
- Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 632–658, Seattle, United States. Association for Computational Linguistics.

- Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. 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.
- Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.
- Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2021. Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations. arXiv:2112.06204 [cs]. ArXiv: 2112.06204.
- Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online. Association for Computational Linguistics.
- Yi Zhang, Sujay Kumar Jauhar, Julia Kiseleva, Ryen White, and Dan Roth. 2021. Learning to decompose and organize complex tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2726–2735, Online. Association for Computational Linguistics.

- Xinyan Zhao and V.G.Vinod Vydiswaran. 2021. Lirex: Augmenting language inference with relevant explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14532–14539.
- 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.