# Agree, Disagree, Explain: Decomposing Human Label Variation in NLI through the Lens of Explanations

Anonymous ACL submission


## Abstract

Natural Language Inference (NLI) datasets often exhibit human label variation. To better understand these variations, explanation-based approaches analyze the underlying reasoning behind annotators' decisions. One such approach is the LITEX taxonomy, which categorizes free-text explanations in English into reasoning types. However, previous work applying such taxonomies has focused on within-label variation: cases where annotators agree on the NLI label but provide different explanations. This paper broadens the scope by examining how annotators may diverge not only in the reasoning type but also in the labeling step. We use explanations as a lens to decompose the reasoning process underlying NLI annotation and to analyze individual differences. We apply LITEX to two NLI datasets and align annotation variation from multiple aspects: NLI label agreement, explanation similarity, and taxonomy agreement, with an additional compounding factor of annotators' selection bias. We observe instances where annotators disagree on the label but provide similar explanations, suggesting that surface-level disagreement may mask underlying agreement in interpretation. Moreover, our analysis reveals individual preferences in explanation strategies and label choices. These findings highlight that agreement in reasoning types better reflects the semantic similarity of explanations than label agreement alone. Our findings underscore the richness of reasoning-based explanations and the need for caution in treating labels as ground truth.


## 1 Introduction

Natural Language Inference (NLI) has long served as a benchmark for language understanding (Dagan et al., 2005; Condoravdi et al., 2003), yet annotation divergence, both in labels and in underlying reasoning, has long been a recognized challenge. Recent work has provided new interpretations of this phenomenon, drawing on insights from linguistics, pragmatics, and conceptual framing (Plank, 2022; Jiang et al., 2023; Kalouli et al., 2023).

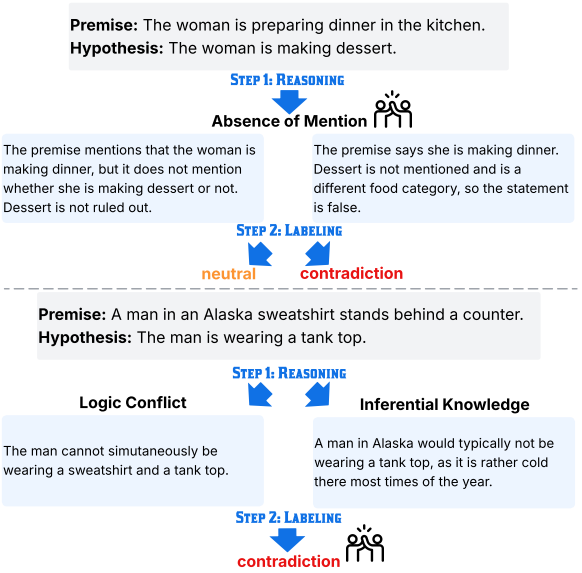


Figure 1: Decomposing NLI annotations into a two-step decision-making framework: Step 1 (Reasoning) and Step 2 (Labeling). The two examples illustrate how annotators may diverge at one of the two steps, featuring the LITEX categories *Logical Conflict* and *Absence of Mention*.

Explanations written in free-text provide valuable insights into the reasoning process behind label assignment (Jiang et al., 2023; Tan, 2022). Instead of treating NLI labels as isolated outcomes, explanations reveal underlying reasoning strategies that annotators employ. Building on Jiang and de Marneffe (2022), which focuses on categorizing linguistic sources of disagreement in the premise–hypothesis pair, Hong et al. (2025) introduced a taxonomy that categorizes explanations according to eight distinct reasoning strategies.

Such a taxonomy enables us to decompose the reasoning process underlying NLI labeling into a two-step decision-making process: (i) **Step 1 (Reasoning)** — after reading the premise and hy-

pothesis, annotators form a logical judgment about their relationship, which is often reflected in the free-text explanation; and (ii) **Step 2 (Labeling)** — based on this rationale, they then select the corresponding NLI label (entailment, neutral, or contradiction). We treat this two-step structure as a working assumption that guides our analysis.

We illustrate both cases in Figure 1 using the LITEX explanation categories. The top example shows a scenario where both annotators use the *Absence of Mention* rationale, but disagree on the label (neutral vs. contradiction), highlighting divergence at the **Labeling** step. The bottom one shows annotators agree on the contradiction label while offering explanations grounded in different categories, reflecting **Reasoning** step variation.

While LITEX is originally designed to study within-label variation (like the second example in Figure 1, which provides different explanations for the same NLI label), it remains unclear how well this approach generalizes to label variation settings, and ***how it can support a deeper understanding of the two-step decision-making process in annotation***. To better understand how and why annotators agree or disagree in NLI, this work decomposes label variation with the help of free-text explanations. By analyzing the interaction of label choices, explanation categories, and explanation text similarities, we offer a deeper view into the reasoning behind disagreement at the dataset and annotator levels. Specifically, our key contributions include:

**From within-label to label variation** We annotate two NLI datasets, LiveNLI (Jiang et al., 2023) and VariErr (Weber-Genzel et al., 2024) with the LITEX taxonomy, enabling the study of reasoning types across both within-label and label variation.

**Annotator tracking via explanation categories** By combining NLI labels with explanation categories, we track individual annotators' reasoning patterns and uncover behavioral consistencies that are not apparent from label distributions alone.

**Quantitative analysis of agreement beyond labels** We measure agreement at three levels: NLI label, explanation category, and text similarity of explanations. Our results show that alignment in reasoning types better correlates with explanation similarity than label agreement alone, emphasizing the importance of explanations for understanding annotator disagreement.

## 2 Related Work

Most benchmark NLI datasets provide multiple annotations per instance, enabling the study of annotation variation. For example, the Stanford NLI (SNLI; Bowman et al. 2015) and MultiNLI (Williams et al., 2018) datasets collect multiple crowd-sourced judgments for each premise–hypothesis pair, which has enabled analyses of systematic disagreement and label variation (Pavlick and Kwiatkowski, 2019; Kalouli et al., 2023). To further address this, adversarial datasets such as ANLI (Nie et al., 2020a) and ambiguity-focused resources like ChaosNLI (Nie et al., 2020b) and AmbiEnt (Liu et al., 2023) were introduced.

Closer to our line of work, recent datasets use free-text explanations and highlights to reveal the reasoning behind NLI labels (Camburu et al., 2018; Jiang et al., 2023; Weber-Genzel et al., 2024). **e-SNLI** expanded SNLI (Bowman et al., 2015) by crowd-sourcing highlight and explanation annotations on the pre-annotated majority label. **LiveNLI** (Jiang et al., 2023) recruited crowd-workers to annotate NLI labels while also providing highlights and explanations (i.e., ecologically valid explanations, produced jointly with the label rather than post hoc). **VariErr** (Weber-Genzel et al., 2024) uses such ecologically valid explanations as a foundation for error detection. Hong et al. (2025) built a taxonomy, **LITEX**, to categorize these free-text explanations, but its scope was limited to e-SNLI and centered on analyzing within-label variation.

A parallel line of work has examined *annotator disagreement*. For example, de Marneffe et al. (2012) and Uma et al. (2022) identified structured patterns of disagreement. For NLI, understanding human reasoning is crucial to interpreting agreement or disagreement. Annotators often rely on various reasoning strategies, such as substitution, negation, bridging inferences, and world knowledge inference (Jiang and de Marneffe, 2022; Kalouli et al., 2023; Sanyal et al., 2024; Hong et al., 2025).

Prior work on annotator decisions often assumes a *two-step* process: interpret the premise–hypothesis pair, then assign an NLI label. However, the interaction between these steps remains underexplored—for instance, when similar reasoning yields different labels or the same label reflects distinct rationales. We address this gap by analyzing this two-step process with a linguistic taxonomy of explanations to clarify how reasoning and labeling interact in NLI.

## 3 From Within-label Variation to Label Variation

To characterize the two-step decision-making process from a global statistical distributional perspective, we use the LITEX taxonomy (Hong et al., 2025) and apply it to two English NLI datasets exhibiting label variation. We then examine the co-occurrence patterns between NLI labels and explanation categories across datasets.

### 3.1 LITEX: a Linguistic Taxonomy of Explanations

LITEX categorizes NLI explanations into two reasoning types: *Text-Based (TB)* and *World-Knowledge (WK)*. TB draws on linguistic evidence in the premise and hypothesis and comprises six subtypes: *Coreference*, *Syntactic*, *Semantic*, *Pragmatic*, *Absence of Mention*, and *Logic Conflict*. WK invokes background knowledge beyond text, covering *Factual Knowledge* and *Inferential Knowledge*. For detailed definitions and annotated examples of all categories, please refer to Hong et al. (2025).

LITEX was originally developed on free-text explanations in e-SNLI (Camburu et al., 2018) to characterize within-label variation, where annotators reach the same label via different rationales (Jiang et al., 2023; Hong et al., 2025). In this paper, we use LITEX beyond its original scope by applying it to additional datasets, examining reasoning strategies not only within-label but also across labels. Furthermore, we use the taxonomy to analyze annotator behavior and to deepen our understanding of the relationship between explanations and NLI labels.

### 3.2 Annotation on LiveNLI and VariErr

To study how annotators diverge in both reasoning and label selection, we apply LITEX to two English datasets with label variation and explanation annotations — **LiveNLI** (Jiang et al., 2023) and **VariErr** (Weber-Genzel et al., 2024).

LiveNLI is a high-quality explanation dataset derived from a subset of MNLI (Williams et al., 2018), containing 122 NLI examples. Each example is annotated by at least 10 crowdworkers, who independently assign one or more NLI labels (*true*, *either*, *false*), highlight relevant spans, and provide free-text explanations (Jiang et al., 2023). For the consistency of our analyses and to align with e-SNLI and VariErr, we map (*true*, *either*, *false*) to (*entailment*, *neutral*, *contradiction*) in this paper.

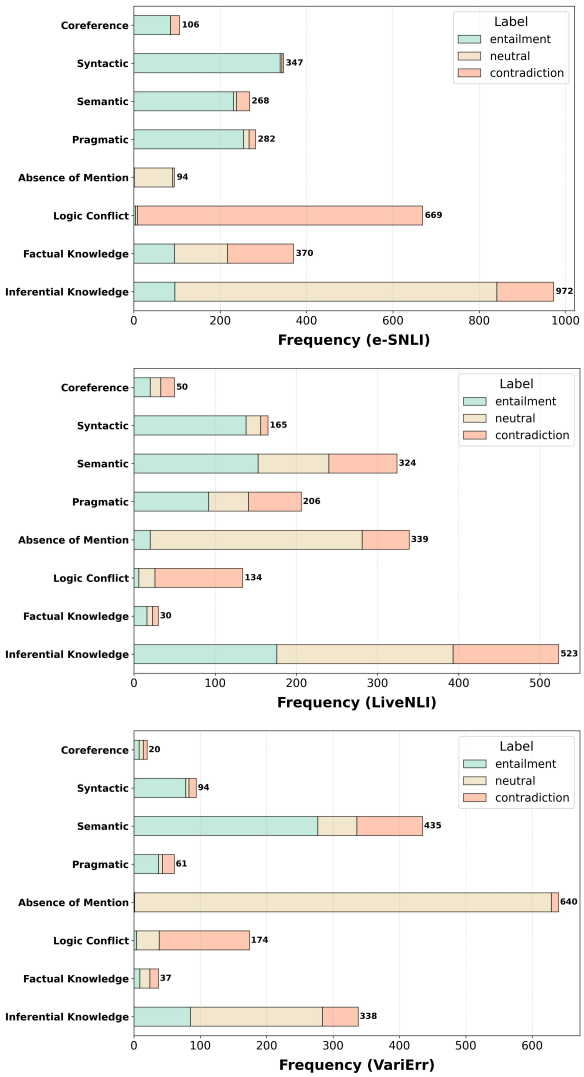


Figure 2: Co-occurrence of LITEX explanation categories and NLI labels across three datasets (e-SNLI, LiveNLI, and VariErr).

VariErr complements LiveNLI by focusing on variation and errors in English NLI. It consists of 1,933 model-generated explanations for 500 re-annotated MNLI items, along with 7,732 human validity judgments (Weber-Genzel et al., 2024). LiveNLI captures natural annotator disagreement, and VariErr introduces both plausible alternative explanations and annotation errors.

A key difference from e-SNLI lies in the labeling scheme. While e-SNLI assigns a single gold label, both LiveNLI and VariErr allow multiple plausible labels per instance and are ecologically valid, with labels and explanations annotated by the same people. These features allow us to analyze the two-step decision-making process in NLI annotation. Appendix A presents a detailed analysis of label distribution per NLI item in the two datasets.

We apply LITEX to annotate explanations from both datasets, 1,404 from LiveNLI and 1,933 from VariErr. All annotations are carried out by a trained annotator.[1] To measure inter-annotator agreement (IAA), we recruited a second annotator to annotate 100 explanations from each dataset independently. We achieved 0.828 on Cohen's Kappa ($\kappa$) for LiveNLI and 0.792 for VariErr, similar to the IAA on e-SNLI (Hong et al., 2025).

### 3.3 LITEX Categories across NLI Labels

We examine how LITEX categories are distributed across NLI labels in the three datasets: e-SNLI, LiveNLI and VariErr in Figure 2. Some categories, such as *Coreference*, display similar trends, consistently appearing less frequently in all three datasets. At the same time, notable differences can also be observed across datasets. For example, *Inferential Knowledge* is the dominant category in both e-SNLI and LiveNLI but is less prominent in VariErr. In contrast, *Absence of Mention* is the most frequent category in VariErr, ranks second in LiveNLI, and occurs relatively less often in e-SNLI.

Additional interesting patterns arise in the **co-occurrences** between NLI labels and taxonomy distributions. The *neutral* label dominates the *Absence of Mention* category across all three datasets. This is consistent with the nature of this reasoning category, which centers on information gaps between the premise and hypothesis. For *Factual Knowledge* and *Inferential Knowledge*, explanations are distributed relatively evenly across the three NLI labels. This reflects the fact that these categories involve the introduction of external or inferred knowledge, without a strong bias toward a specific label. The *Syntactic*, *Semantic*, and *Pragmatic* categories are more strongly associated with entailment (despite slight variations across datasets) suggesting that annotators often rely on evidence from different linguistic levels within the premise and hypothesis to establish entailment relationships.

In sum, although the absolute distributions of explanation categories differ across datasets, the label distribution and dominant label for each category remain highly consistent. The observed differences in category distribution may stem from factors such as annotator backgrounds and preferences, dataset-specific item selection, and annotation guidelines. Nevertheless, the stable category–label co-occurrence patterns indicate that, from a global distributional perspective, LITEX provides **a reliable characterization of reasoning types that generalize across explanations** in different NLI datasets.

[1] Both annotators are students trained with the taxonomy definitions and examples, and were paid according to the national standard.

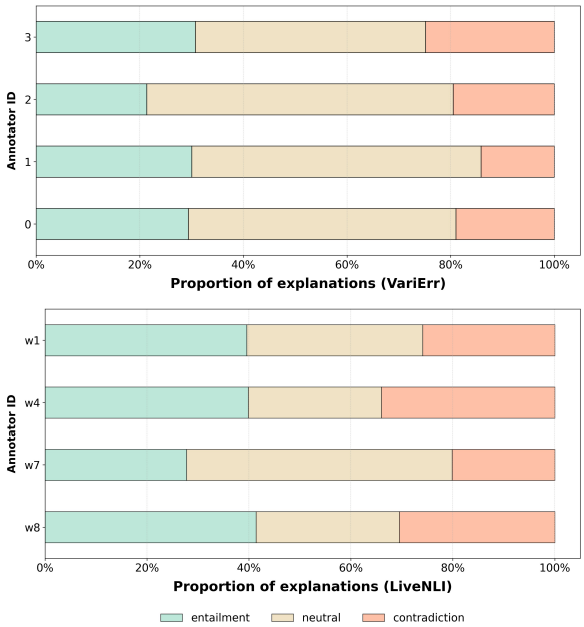


Figure 3: Distribution of NLI labels (entailment, neutral, contradiction) across LiveNLI and VariErr annotators. The legend at the bottom specifies the color–label correspondence, while the area of each color segment represents the number of instances assigned to that label.

## 4 Label and Reasoning Preferences among Individual Annotators

To gain a deeper understanding of disagreement in NLI, we propose to jointly analyze annotator preferences over both NLI labels and explanation categories on LiveNLI and VariErr. Specifically, we track individual annotators across two dimensions: their **NLI label preferences** (e.g., tendency to overuse "neutral"), and their **reasoning-type preferences**, manual taxonomy classification of their free-text explanations annotated in this paper.

Unlike VariErr, which is annotated by the same four annotators, LiveNLI involves a much larger annotator pool (Jiang et al., 2023). As a result, each annotator in LiveNLI only annotated a small subset of data. To align with the setup in VariErr, we chose a group of four annotators {`w1`, `w4`, `w7`, `w8`} from LiveNLI for comparison based on having the highest number of overlapping annotated items (115 in total). Appendix B provides detailed tables of NLI label distributions and explanation category

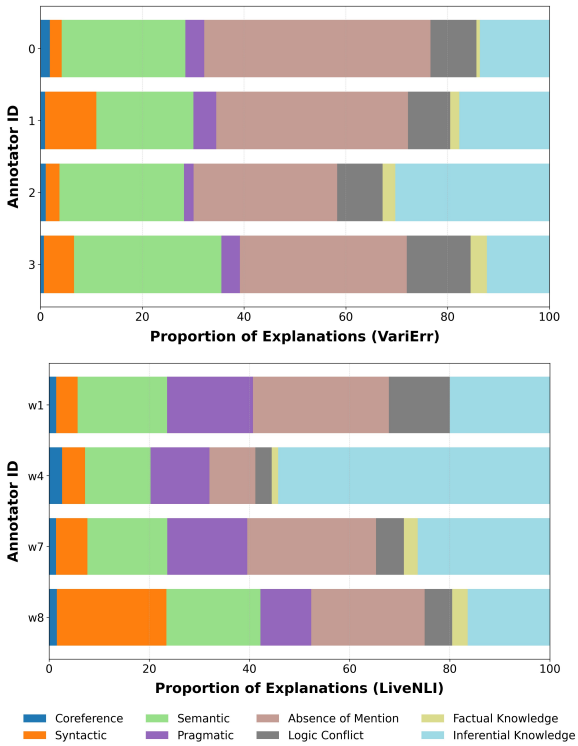


Figure 4: Distribution of explanation category per annotators in LiveNLI and VariErr. Colors correspond to different explanation categories.

preferences for the four annotators in LiveNLI and VariErr, computed over the intersection of items they jointly annotated in each dataset.

**Label Preferences** Figure 3 illustrates the distribution of NLI labels assigned by four annotators in both datasets. For VariErr, while all annotators exhibit a noticeable preference for the *neutral* label, the degree of this preference varies. Annotators `0`, `1`, and `2` all assign *neutral* in over 50% of their examples, with Annotator `2` showing the strongest bias—nearly 60% of their annotations are *neutral*. In contrast, Annotator `3` demonstrates a more balanced labeling behavior, with a more even spread across *entailment*, *neutral*, and *contradiction*, resulting in the lowest *neutral* proportion (44.42%).

As for LiveNLI, three annotators (`w1`, `w4`, `w8`) show a slight preference toward the *entailment* label, each assigning it to approximately 40% of examples. In contrast, annotator `w7` exhibits a stronger preference for the *neutral* label, assigning it in over half of the cases (52.08%). The *contradiction* labels are the least common overall. Compared to annotators in VariErr, where *neutral* was the dominant label for all four annotators, the LiveNLI group shows more varied labeling tendencies.

**Reasoning-Type Preferences** Figure 4 presents the distribution of explanation categories of the annotators. Individual differences emerge in how annotators ground their inferences. For VariErr, annotator `0` shows a dominant reliance on *Absence of Mention* (44.44%) and *Semantic* reasoning (24.31%), with only minimal use of world knowledge-based types such as *Factual Knowledge* (0.69%) and *Inferential Knowledge* (13.66%). This pattern suggests a preference for surface-level paraphrastic inference, rather than deeper reasoning. In contrast, Annotator `1` exhibits a more balanced distribution, with moderate use of the reasoning strategies. Annotator `2` stands out with a very strong emphasis on *Inferential Knowledge* (30.29%), while still relying on *Semantic* explanations (24.48%). This suggests a knowledge-intensive reasoning, grounded in world knowledge and inferencing beyond what is stated. Similarly, Annotator `3` relies more on knowledge-based reasoning, using *Inferential Knowledge* (12.3%) and *Factual Knowledge* (3.2%) more frequently.

For LiveNLI, several trends emerge from the distribution. First, *Semantic* and *Absence of Mention* explanations are consistently among the most frequently used categories across annotators, suggesting that both lexical-semantic inferences and missing information play a central role. Second, we observe notable variation in the use of *Inferential Knowledge*: annotator `w4` relies on this category in over half of their explanations (54.25%), while `w1` and `w8` use it far less frequently (20.00% and 16.41% respectively), indicating divergent preferences in relying on external world knowledge. Similarly, *Syntactic* explanations are prominent for `w8` (21.88%) compared to the others, reflecting a possible inclination toward structural reasoning. Conversely, *Logical Conflict*, *Factual Knowledge*, and *Coreference* are relatively rare across annotators, suggesting these reasoning types are less frequently invoked or salient in this LiveNLI subset.

Overall, the annotator-level analysis facilitated by LITEX over the two datasets reveals that **different annotators tend to adopt distinct reasoning strategies**—arriving at different NLI labels for the same premise–hypothesis pairs. Observing only the distribution of NLI labels is insufficient to uncover the underlying reasoning rationales. To gain a deeper understanding of annotator behavior, we next conduct a fine-grained item-level analysis that disentangles variation in reasoning from variation in label assignment.

| Agreement Class | Entropy | Support (%) | Category Agreement | token 1-gram | token 2-gram | POS 1-gram | POS 2-gram | cosine (%) | euclidean (%) |
|---|---|---|---|---|---|---|---|---|---|
| *VariErr* | | | | | | | | | |
| Full (4-0-0) | 0.00 | 43.75 | **0.76** | **35.05** | **11.53** | 74.21 | **35.06** | **52.87** | 51.89 |
| Partial (3-1-0) | 0.81 | 28.95 | 0.60 | 34.72 | 10.62 | **78.31** | 34.85 | 52.81 | **51.96** |
| Two Pairs (2-2-0) | 1.00 | 23.36 | 0.56 | 30.80 | 8.50 | 73.47 | 31.23 | 49.22 | 51.02 |
| Divergent (2-1-1) | 1.50 | 3.95 | 0.50 | 32.02 | 9.96 | 70.37 | 31.50 | 48.21 | 50.91 |
| *LiveNLI* | | | | | | | | | |
| Full (4-0-0) | 0.00 | 21.74 | **0.62** | 40.31 | 10.26 | **88.89** | 41.96 | **58.05** | **53.24** |
| Partial (3-1-0) | 0.81 | 34.78 | 0.56 | **40.44** | **11.95** | 86.99 | **44.02** | 54.47 | 52.27 |
| Two Pairs (2-2-0) | 1.00 | 23.48 | 0.60 | 38.97 | 10.67 | 88.38 | 43.24 | 55.33 | 52.84 |
| Divergent (2-1-1) | 1.50 | 20.00 | 0.54 | 36.61 | 8.99 | 85.09 | 41.44 | 53.63 | 52.35 |

Table 1: Aggregated statistics across agreement classes for LiveNLI and VariErr, based on how many annotators agree on the NLI label. We report the label entropy, the percentages of support items, corresponding category agreement, and average pairwise explanation similarity. Color coding highlights relative deviations within each dataset: cells shaded in blue or red indicate values that are higher or lower than expected given the level of label agreement. Darker shades correspond to larger deviations in ranking, and lighter shades indicate smaller deviations.

## 5 Measuring and Interpreting Agreement

Building on earlier observations that NLI distributions are insufficient to uncover reasoning rationales of individual annotators, this section takes a closer look at how to measure and interpret agreement and disagreement in NLI tasks *at the instance level*. We first quantify annotator agreement across three dimensions: NLI labels, explanation categories, and textual similarity between explanations. We then compute pairwise agreement and visualize it using conditional probability heatmaps to examine the relationship between reasoning and labeling. At last, we present a case study from both datasets to illustrate how annotators may align or diverge in their label decisions and reasoning processes.

### 5.1 Quantifying Annotator Agreement Beyond Labels

To gain a clearer picture of annotator agreement in NLI labels and LITEX categories, we group NLI items based on the set of annotators who labeled them and their corresponding agreement scenario. We define four NLI **label agreement** classes based on how many of the four annotators agree: *Full Agreement* (4-0-0) indicates all four annotators assigned the same label; *Partial Agreement* (3-1-0) refers to cases where three annotators agreed while one differed; *Two Pairs Agreement* (2-2-0) denotes a balanced split, with two annotators agreeing on one label and the other two on a different label; and *Divergent* (2-1-1) captures maximal disagreement, where three different labels are assigned.

**Category agreement** is measured via the average Jaccard similarity between the explanation categories. Since each explanation has only one category, this reduces to computing the proportion of explanation pairs that share the same category:

$$\text{Jaccard}(a, b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

We follow Chen et al. (2025); Hong et al. (2025) to quantify **textual similarities between explanations** using measures from Giulianelli et al. (2023). Lexical and syntactic similarities evaluate overlapping unigrams and bigrams on tokens and POS tags. We use cosine and Euclidean to measure semantic similarity between sentence embeddings, obtained using the `all-distilroberta-v1`[2] model from SentenceTransformers (Sanh et al., 2019). Scores are averaged pairwise across four explanations.

Table 1 summarizes the label entropy, number of supporting items, category agreement, and average explanation similarity for each label agreement scenario. Examining across evaluation metrics, the color coding highlights deviations in ranking across agreement classes relative to label agreement, though the absolute differences remain small.

VariErr generally exhibits less ranking deviation than LiveNLI, with a matched ranking between label agreement class and category agreement, and only light deviations in textual similarity measures. Full agreement is also considerably more frequent in VariErr (43.75%) than in LiveNLI (21.74%), indicating that annotators tend to reach agreement more easily in VariErr. Moreover, all ranking deviations in VariErr concern an additional NLI label:

[2] https://huggingface.co/sentence-transformers/all-distilroberta-v1

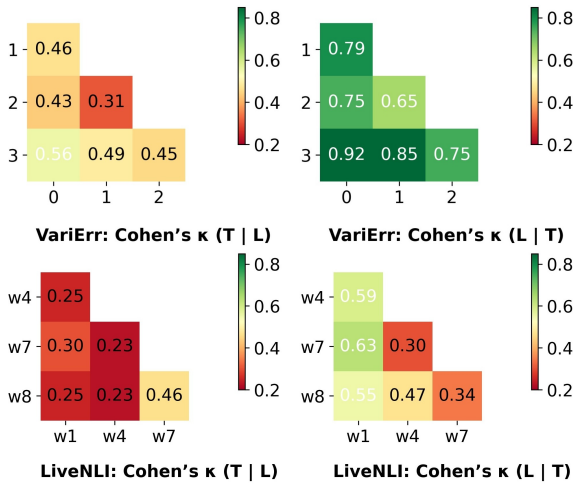


Figure 5: Pairwise annotator agreement (conditional Cohen's $\kappa$) between taxonomy matches (T) and label matches (L).

between full and partial (4-0-0 vs. 3-**1**-0) and between two pairs and divergent (2-2-0 vs. 2-1-**1**), whereas many LiveNLI ranking deviations stem from different distributions of the same labels, i.e., between partial and two pairs (**3**-**1**-0 vs. **2**-**2**-0).

Looking at the text similarity measures, we found that cosine similarity resonates the most with label agreement on both datasets, exhibiting moderate differences in scores across classes and only minor deviations in ranking on LiveNLI between partial and two pairs. More interestingly, the pattern of cosine similarity aligns more closely with category agreement than label agreement. This observation **tentatively suggests that shared reasoning types may better capture the semantic similarity of explanations than label agreement**.

### 5.2 Pairwise Annotator Agreement on Reasoning and Labeling

To quantify how annotators align their reasoning processes with labeling decisions, we examine pairwise agreement patterns between explanation categories (reasoning) and NLI labels (labeling) using conditional Cohen's $\kappa$ scores (Figure 5). This analysis is grounded in the two-step annotation process as in Figure 1, where annotators first engage in reasoning and then select a label based on that reasoning, allowing us to examine variation at each step. Specifically, for each pair of annotators, we compute two conditional $\kappa$ scores:

$$\kappa(T \mid L), \tag{2}$$

which measures taxonomy agreement (reasoning alignment) on the subset of instances where the two annotators agree on the NLI label, and

$$\kappa(L \mid T). \tag{3}$$

which measures label agreement on the subset of instances where the two annotators match in their reasoning categories. Unlike raw agreement, Cohen's $\kappa$ adjusts for chance, providing a more robust measure of inter-annotator alignment.

Like earlier findings, VariErr shows higher agreement than LiveNLI under both conditional $\kappa$ metric, indicating more stable reasoning–label mappings in VariErr, whereas LiveNLI reflects greater variation in annotator interpretations. Across both datasets, annotators show higher $\kappa$ for labels conditioned on taxonomy matches $\kappa(L \mid T)$ than for taxonomy conditioned on label matches $\kappa(T \mid L)$. Our two-step annotation process can explain the asymmetry between the two conditional probabilities. When annotators share the same taxonomy categories, they are highly likely to assign the same label, indicating that divergence in the second step (labeling) exists but occurs less frequently. In contrast, when annotators agree on the final label, their reasoning categories often diverge, suggesting that **divergence in the reasoning step is relatively more common than that in the labeling step**. Therefore, to capture label variation more accurately and informatively, it is crucial to focus on the reasoning step, as it is the dominant source of variation.

### 5.3 Case Study: Diverging Interpretations on the Same NLI Instance from Two Datasets

To conclude our analysis, we present a case study that illustrates how annotators from two different datasets interpret the same NLI instance in divergent ways, both in the reasoning step and the labeling step. Table 2 presents one example annotated by the eight annotators we analyzed in earlier sections, four from LiveNLI and four from VariErr. For additional qualitative illustrations, further examples are provided in the Appendix C.

Comparing the two datasets, we observe that while the set of chosen NLI labels (neutral and contradiction) and explanation categories (*Absence of Mention* and *Inferential Knowledge*) is the same, the distribution of these choices differs: in LiveNLI, three annotators opted for contradiction, whereas in VariErr, three chose neutral. In terms of explanation categories, LiveNLI annotators are evenly split, with two selecting *Absence of Mention* and two choosing *Inferential Knowledge*. In VariErr, three

**Premise:** The author began with a set of hunches or hypotheses about what can go wrong in agency management, and what would be evidence supporting—or contradicting—these hypotheses.
**Hypothesis:** The hunches provided by the author weren't realistic as it pertains to agency management.
[CHAOSNLI] [E,N,C]: [0.64, 0.06, 0.30]

| Dataset | Annotator | NLI Label | Explanation | Explanation Category |
|---|---|---|---|---|
| LiveNLI | w1 | Neutral | The context notes that the hunches were provided but there is no information on the veracity or plausibility of those hunches, so the statement could be true or false because it is not known if they were realistic based on the information provided. | Absence of Mention |
| | w4 | Contradiction | If an author is planning to write about a certain topic, odds are they have enough knowledge to have informed opinions on the subject right from the beginning. Odds are much higher the statement is false because these informed opinions would likely be at least somewhat realistic. | Inferential Knowledge |
| | w7 | Contradiction | The author gave evidence to support his hunches, so it is unlikely that the hunches were unrealistic. | Inferential Knowledge |
| | w8 | Contradiction | It wasn't stated that the hunches were not realistic. | Absence of Mention |
| VariErr | 0 | Neutral | It is not clear whether the hunches provided by the author were realistic or not. | Absence of Mention |
| | 1 | Neutral | It is not clear how realistic the hypotheses were. | Absence of Mention |
| | 2 | Neutral | The judgment of the hunches is not given in the context. | Absence of Mention |
| | 3 | Contradiction | The hunches could be realistic, as the author provide potential evidence supporting these hypotheses. | Inferential Knowledge |

Table 2: Case study of an NLI item annotated in both LiveNLI and VariErr. Despite identical inputs, annotators diverge in NLI labels and explanation strategies.

annotators attribute their reasoning to *Absence of Mention*, while one goes for *Inferential Knowledge*.

The example further illustrates the two-step reasoning process in NLI. Among LiveNLI annotators, w4, w7, and w8 agree on the NLI label (contradiction), but provide explanations grounded in different reasoning types, namely *Inferential Knowledge* vs. *Absence of Mention*, showcasing within-label variation and divergence at reasoning step. In contrast, w1 and w8 both rely on the *Absence of Mention* rationale, yet arrive at different NLI labels (neutral vs. contradiction), pointing to divergence at the labeling step. Meanwhile, VariErr annotators demonstrate both label and reasoning agreement. Annotators 0, 1, and 2 all classify the instance as neutral, supported by nearly identical explanations and shared categorization as *Absence of Mention*. This coherence suggests alignment at both steps of the reasoning process, from interpretive rationale to label decision. Annotator 3 stands out with Inferential Knowledge reasoning and a contradiction label, further emphasizing how divergent reasoning can lead to different label choices.

Overall, this example shows **how combining NLI labels with explanation categories reveals deeper patterns of disagreement and agreement**—distinctions that would remain hidden if one only considered label distributions alone, while reasoning information further illuminates the underlying sources of variation.

## 6 Conclusion

Understanding why annotators diverge is key to interpreting NLI labels. We extend LITEX to two NLI datasets with free-text explanations and jointly analyze within-label and cross-annotator label variation. By combining labels, taxonomy categories, and explanation texts, we uncover reasoning patterns that label distributions miss. Across analyses, taxonomy categories track explanation-text similarity more closely than labels, emphasizing reasoning paths over surface label agreement.

Our results suggest three broader implications. First, NLI explanations differ in what they ground the decision on, and these differences correlate with how often annotators agree or disagree. Second, agreement in reasoning type is more predictive of label agreement than shared labels are of reasoning-type agreement, indicating that reasoning provides a key trace of variation. Third, annotators exhibit distinct reasoning profiles, which can lead to divergent labels on the same instances. These findings motivate future NLI work to explicitly track reasoning types during dataset construction (e.g., which categories appear and which ones tend to trigger disagreement) and to use reasoning types to stratify evaluation sets, offering a more diagnostic view of where models succeed or fail.

More broadly, labels and even aggregated label distributions can hide how annotators arrive at their interpretations. Complementing labels with explanations provides a clearer window into annotators' inferential steps. While the ideal explanation format remains open, free-text explanations retain a unique advantage in revealing how different interpretations can emerge from the same input.

Future work can extend LITEX to allow multiple reasoning strategies per explanation, capturing more interactions. Modeling annotator backgrounds may further reveal systematic sources of disagreement. Finally, integrating this framework with explanation generation could improve the quality and evaluation of model rationales, and applying it to other domains would test its generality.

## Limitations

Our work has several limitations. First, our analysis relies on the category set defined in LITEX, which may not fully capture the complexity or compositionality of human reasoning; extending the taxonomy to allow multiple categories per explanation is an important direction.

Second, our datasets are substantially smaller than large-scale NLI benchmarks such as SNLI or MNLI. However, the analysis is based on 1,404 explanations from LiveNLI and 1,933 explanations from VariErr, which is relatively large for manually categorized free-text explanations. Accordingly, our goal is not to estimate fine-grained population statistics for all possible annotators, but to uncover robust qualitative and quantitative patterns in how reasoning types relate to label variation.

Third, our annotator-level analysis focuses on annotators with sufficient per-annotator coverage: we analyze four annotators who each contributed explanations for over 100 items on a shared set of instances, yielding substantial within-annotator evidence for comparing reasoning profiles. Nevertheless, including more annotators and more diverse annotator populations would strengthen generalization.

Finally, we do not directly model annotator-pool effects (e.g., background knowledge, fatigue, or instruction framing), and our similarity measures (e.g., sentence-embedding cosine similarity and Jaccard agreement over categories) provide only a partial view of reasoning variation, potentially missing deeper pragmatic or structural divergences.

## Ethical considerations

We do not foresee any ethical concerns associated with this work. All analyses were conducted using publicly available datasets and models. No private or sensitive information was used. Additionally, we will make our code, analysis, and documentation publicly available to ensure transparency and reproducibility.

## Use of AI Assistants

The authors acknowledge the use of ChatGPT for correcting grammatical errors and enhancing the coherence of the final manuscripts.

## References


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.

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.

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10777–10802, Vienna, Austria. Association for Computational Linguistics.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. pages 177–190.

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.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2025. Litex: A linguistic taxonomy of explanations for understanding within-label variation in natural language inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. 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.

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.

Aikaterini-Lida Kalouli, Hai Hu, Alexander F. Webb, Lawrence S. Moss, and Valeria de Paiva. 2023. Curing the SICK and other NLI maladies. *Computational Linguistics*, 49(1):199–243.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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

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.

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.

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. 2024. Are machines better at complex reasoning? unveiling human-machine inference gaps in entailment verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10361–10386, Bangkok, Thailand. 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 Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. 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.

## A Label Distribution per NLI Item in LiveNLI and VariErr

To better understand the distribution characteristics of NLI labels in the selected two datasets, we visualize the aggregated label probabilities for each item.

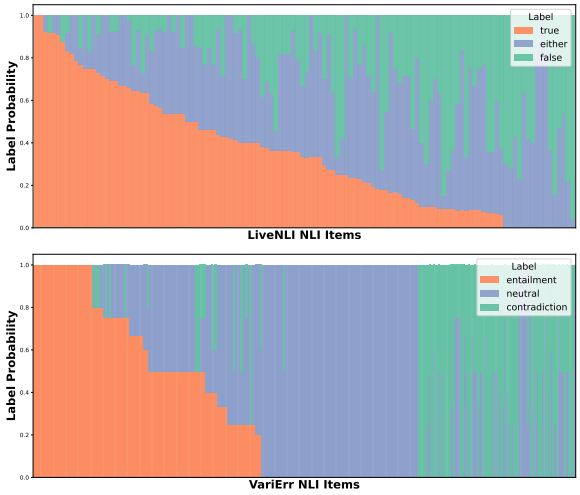


Figure 6: Normalized label distribution per NLI item in LiveNLI and VariErr. Items are sorted by the proportion of *entailment/true* labels.

Figure 6 presents stacked bar charts of the normalized NLI label distributions across all items, sorted by the proportion of *entailment/true* labels to highlight overall patterns of annotator agreement. Both datasets reveal a range of label variation, with many items reflecting ambiguity or disagreement among annotators. However, their distributions differ in characteristic ways. LiveNLI shows greater

diversity, particularly in cases where the *either* label dominates or co-occurs substantially with the other two categories. In contrast, VariErr exhibits more concentrated distributions, with fewer items displaying high levels of ambiguity. Nonetheless, some items still reveal substantial variation in label assignment, pointing to challenging or underspecified NLI cases. These findings motivate our subsequent analysis of how explanation similarity varies across explanation categories and NLI labels.

## B Annotator-wise distribution of NLI labels and explanation categories

Table 3 presents the detailed NLI label distributions and explanation category preferences for the four annotators in LiveNLI and VariErr, computed over the intersection of items they jointly annotated in each dataset. This table provides the full statistics underlying the summary reported in Section 4.

## C Case Study: NLI items annotated in both LiveNLI and VariErr

Table 4 presents more illustrative examples of NLI items that were annotated in both LiveNLI and VariErr by the selected annotator groups. These examples provide additional qualitative context for the agreement patterns discussed in the main text.

| Annotator ID | #Ex | Ent | Neu | Con | Coref | Synt | Sem | Prag | Abs | Logic | FK | IK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *VariErr* | | | | | | | | | | | | |
| 0 | 432 | 29.40 | 51.62 | 18.98 | 1.85 | 2.31 | 24.31 | 3.70 | 44.44 | 9.03 | 0.69 | 13.66 |
| 1 | 446 | 30.04 | 55.83 | 14.13 | 0.90 | 10.09 | 19.06 | 4.48 | 37.67 | 8.30 | 1.79 | 17.71 |
| 2 | 482 | 21.37 | 59.13 | 19.50 | 1.04 | 2.70 | 24.48 | 1.87 | 28.22 | 8.92 | 2.49 | 30.29 |
| 3 | 439 | 30.75 | 44.42 | 24.83 | 0.68 | 5.92 | 28.93 | 3.64 | 32.80 | 12.53 | 3.19 | 12.30 |
| *LiveNLI* | | | | | | | | | | | | |
| w1 | 140 | 39.29 | 34.29 | 25.71 | 1.43 | 4.29 | 17.86 | 17.14 | 27.14 | 12.14 | 0.00 | 20.00 |
| w4 | 153 | 39.87 | 26.14 | 33.99 | 2.61 | 4.58 | 13.07 | 11.76 | 9.15 | 3.27 | 1.31 | 54.25 |
| w7 | 144 | 27.78 | 52.08 | 20.14 | 1.39 | 6.25 | 15.97 | 15.97 | 25.69 | 5.56 | 2.78 | 26.39 |
| w8 | 128 | 41.41 | 28.13 | 30.47 | 1.56 | 21.88 | 18.75 | 10.16 | 22.66 | 5.47 | 3.13 | 16.41 |

Table 3: Annotator-wise distribution of NLI labels (entailment, neutral, contradiction) and explanation categories in LiveNLI and VariErr. Percentages (%) are shown for both types of distributions.

**Premise:** If you have any questions about this report, please contact Henry R. Wray, Senior Associate General Counsel, at (202) 512-8581.
**Hypothesis:** Henry R. Wray can be reached at (555) 512-8581.
[CHAOSNLI] [E,N,C]: [0.21, 0.76, 0.03]

| Dataset | Annotator | NLI Label | Explanation | Explanation Category |
|---|---|---|---|---|
| LiveNLI | w1 | Contradiction | The two phone numbers provided for Wray do not match, thus the statement is likely to be false since it does not present the same number as in the context. | Logic Conflict |
| | w4 | Contradiction | Henry R. Wray's actual phone number is listed in the context. While it's true he could have multiple contact numbers, it's more likely that the context is his only contact number since it's connected to his business. | Inferential Knowledge |
| | w7 | Contradiction | The phone number for Henry Wray is (202) 512-8581, so (555) 512-8581 is probably not going to reach him. | Inferential Knowledge |
| | w8 | Contradiction | Henry R. Wray can be reached at a difference number which is (202) 512-8581. | Semantic |
| VariErr | 0 | Contradiction | The number of Henry R. Wray mentioned in the statement is wrong. | Logic Conflict |
| | 1 | Contradiction | The phone number starts with (202) not with (555). | Logic Conflict |
| | 2 | Contradiction | The number is wrong, it should be (202) 512-8581 not (555) 512-8581. | Logic Conflict |
| | 3 | Contradiction | wrong phone numbers. | Logic Conflict |

**Premise:** It is, as you see, highly magnified.
**Hypothesis:** It is plain for you to see that it is amplified.
[CHAOSNLI] [E,N,C]: [0.76, 0.10, 0.14]

| | | | | |
|---|---|---|---|---|
| LiveNLI | w1 | Entailment | Magnification is a form of amplification. Therefore, if something is easily seen as magnified, it follows that it's plain to see it's amplified. | Inferential Knowledge |
| | w4 | Entailment | Whatever the speaker is pointing out is clearly magnified or "amplified", and the person they are showing this to is merely having this pointed out to them, making the statement most likely true. | Inferential Knowledge |
| | w7 | Entailment | The words magnified and amplified have the same basic meaning, so the statement agrees with the context. | Semantic |
| | w8 | Entailment | Both amplified and magnified mean the same thing so the statement is true. | Semantic |
| VariErr | 0 | Entailment | Both the context and the statement suggest that it is magnified. | Semantic |
| | 1 | Entailment | The statement is a paraphrase of the context. | Syntactic |
| | 2 | Entailment | ẗighly magnifiedc̈an be interpreted ämplified. | Semantic |
| | 3 | Entailment | It can be seen, and it is magnified. | Semantic |

**Premise:** A clean, wholesome-looking woman opened it.
**Hypothesis:** The woman was trying to be desecrate.
[CHAOSNLI] [E,N,C]: [0.68, 0.31, 0.01]

| | | | | |
|---|---|---|---|---|
| LiveNLI | w1 | Contradiction | The context notes that the woman is clean and wholesome-looking while the statement notes that the woman was being disrespectful, which is not compatible. Thus, it is likely to be false. | Semantic |
| | w4 | Neutral | The statement is nonsensical. Hence there's no information in it, either true or false, to be compared to the context. | Absence of Mention |
| | w7 | Neutral | Just because the woman was wholesome-looking does not mean that she was acting in a discreet manner. | Inferential Knowledge |
| | w8 | Contradiction | The woman was described as wholesome and wouldn't desecrate something. | Semantic |
| VariErr | 0 | Neutral | The context doesn't mention anything about desecration. | Absence of Mention |
| | 1 | Neutral | It's not clear what the woman was trying to be. | Absence of Mention |
| | 2 | Neutral | The attempt of the woman is not given in the context. | Absence of Mention |
| | 3 | Contradiction | Context is a compliment, statement is a negative comment. | Logic Conflict |

Table 4: Examples of NLI items annotated in both LiveNLI and VariErr by the selected annotator groups (w1, w4, w7, w8 for LiveNLI; 0, 1, 2, 3 for VariErr).