# Reference *(In-)*Determinacy in Natural Language Inference

Anonymous ACL submission

## Abstract

Natural Language Inference (NLI) provides a general task format for evaluating the semantic relations between two pieces of text, which can be useful for various applications such as fact verification and text attribution. However, existing datasets for NLI and models trained on these datasets make assumptions about the context from which the premise and hypothesis are sampled. In this paper, we revisit this *reference determinacy* (RD) assumption in NLI, i.e., the premise and hypothesis are assumed to refer to the same context when human raters annotate a label. While RD is a practical assumption for constructing a new NLI dataset, we observe that current NLI models—which are typically trained solely on hypothesis-premise pairs created with the RD assumption—fail in many practical settings in which the premise and hypothesis may refer to different contexts. To highlight the impact of this phenomenon in real-world use cases, we introduce the REFNLI, a diagnostic benchmark for identifying reference ambiguity in NLI examples. In REFNLI, the premise is retrieved from a knowledge source (i.e. Wikipedia) and does not necessarily refer to the same context as the hypothesis. With REFNLI[1], we demonstrate that finetuned NLI models and few-shot prompted LLMs both fail to recognize context mismatch, leading to $> 80\%$ false contradiction and $> 50\%$ entailment predictions. We discover that the existence of reference ambiguity in NLI examples can in part explain the inherent human disagreements in NLI, and provide insight into how the RD assumption impacts NLI dataset creation process.

## 1 Introduction

Natural Language Inference (NLI), or Recognizing Textual Entailment (RTE), provides a general task format for evaluating the semantic relation between two pieces of text, where a system is expected to predict if a hypothesis statement can be inferred from a given premise. For the past few decades, NLI has been the centerpiece for the development and evaluation of language understanding systems (Dagan et al., 2005; Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020a).

**Premise**: A black race car starts up in front of a crowd of people.
**Hypothesis**: A man is driving down a lonely road.

**Label** (Assuming *Reference Determinacy*): Contradiction (5/5)
**Label** (without *Reference Determinacy*): Neutral

Table 1: An example from the SNLI dataset (Bowman et al., 2015) with all five annotators agreeing on the hypothesis contradicting the premise, under the *reference determinacy* assumption, i.e. the events described in the premise and hypothesis happen on the same road. Without the assumption, the label would likely be neutral.

As the use of NLI now spreads across a wider variety of downstream applications, such as text classification (Yin et al., 2019), fact verification (Schuster et al., 2021), hallucination detection (Kryscinski et al., 2020), text attribution (Gao et al., 2023), etc., it is important to understand how the *definitions* and *assumptions* made for collection of previous NLI datasets and models trained on them affect their usefulness in downstream use cases.

In this paper, we revisit and study the effect of *reference determinacy* (RD), a common assumption formed in the labeling of NLI datasets. With RD, the NLI label between a pair of premise and hypothesis is annotated under the assumption that the pair refer to the same context (Bowman et al., 2015). We illustrate the idea behind RD through an example in Table 1, where the premise and the hypothesis describe two different events. The premise *contradicts* the hypothesis (i.e., premise $\rightarrow \neg$ hypothesis) only when we opt to assume that the two events happen on the same road at the same time. Otherwise, the pair would be labeled *neutral*, as the two events are most likely unrelated.

[1] The REFNLI benchmark will be released under CC-BY-SA 4.0

RD is a practical assumption for the NLI label definition. Without the RD assumption, the entailment and contradiction relations would only exist when the hypothesis and premise describe functional relations that are universally true or false (Ritter et al., 2008), e.g. factual knowledge about an entity. For this reason, most large-scale NLI benchmarks follow the RD assumption during their annotation processes (§2.2). However, if we train NLI models exclusively on hypothesis-premise pairs created with the RD assumption, this could lead to the resulting models having limited ability to recognize if a hypothesis is relevant to a premise.

We demonstrate the trickle-down effects of such NLI model behavior in downstream tasks such as fact verification. Specifically, we sample claims from FEVER (Thorne et al., 2018) and VitaminC (Schuster et al., 2021) and study how NLI models behave when used to verify against evidence retrieved from the web. From the sampled claims, we construct the RefNLI benchmark (§3), which features 1,143 NLI pairs with expert judgements for whether the premise and hypothesis refer to the same context, as well as the correct NLI label.

With RefNLI, we observe that both finetuned NLI models as well as LLMs few-shot prompted to classify 3-way NLI labels often fail to recognize context mismatches, which leads to many false entailment and contradiction predictions. On five popular NLI datasets (§4), we demonstrate that different combinations of training datasets result in similar type of reference (in-)determinacy problem in the finetuned model. This indicates the existence of a reference determinacy bias in all five datasets, which we discuss in the context of how each of the five datasets are created. We propose strategies to filter out entailment or contradiction examples labeled only due to the reference determinacy assumption, and show this can mitigate the reference determinacy bias of finetuned NLI models at inference time.

Reference determinacy, we discover, can also partly explain part the distribution of human disagreements of NLI labels, a problem known to be widespread in popular NLI datasets (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b). Our analysis shows that human typically disagree more on examples where reference determinacy cannot be safely assumed, and disagreements happen when annotators are instructed to do so regardless.

In summary, our contributions in the paper are:

- We introduce the RefNLI benchmark, a dataset featuring 1,143 examples for studying the the effect of reference determinacy in NLI, a common assumption in the creation processes of NLI datasets.
- With RefNLI, we investigate the downstream impact of the reference determinacy assumption of NLI dataset creation process. We show that finetuned NLI models and LLMs exhibit reference determinacy bias and often fail to recognize context mismatches.
- We discover and study the connection of the reference determinacy assumption to the inherent human disagreement on NLI labels.

## 2 Preliminaries

### 2.1 Textual Entailment and Contradiction

*Textual Entailment* (Dagan and Glickman, 2004) is defined as a directional relation between a pair of text expressions, namely a premise and a hypothesis. We say that the premise entails the hypothesis if humans would typically infer that the hypothesis is most likely true in the context of the premise. Similarly, we say that the premise contradicts the hypothesis if the hypothesis is highly unlikely to be true given the information described in premise (de Marneffe et al., 2008).

The task of NLI or RTE is usually formatted as a three-way classification of a premise and a hypothesis into *entailment*, *contradiction* or *neutral*, where *neutral* indicates that the premise neither entails or contradicts the claims in full. NLI has widely been adopted as a general task format in the NLP community (Wang et al., 2018, 2019), and many datasets have since been created to facilitate the evaluation of model's language learning capabilities (Poliak, 2020).

### 2.2 The Reference Determinacy Assumption

When we create and label NLI examples, *reference determinacy* (RD) is a practical assumption for guaranteeing the correctness and consistency of annotated labels. For instance, suppose a hypothesis and premise pair both mention *John Doe*, the perceived entailment or contradiction relation could change based on whether we believe the two Johns Doe are a single real-world person.

**The creation processes of most NLI datasets assume reference determinacy.** For example, in SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), annotators were asked to write novel hypotheses that are either true/false/neutral in the context of a given premise. During labeling, the hypothesis is interpreted in the context of the premise, where entities and events in the two are assumed to be co-refer between the hypothesis and premise As a result, we see examples like in Table 1, where majority of the annotators would agree on the contradiction or entailment label, when the premise and hypothesis likely refer to different events without the RD assumption.

Following MNLI and SNLI, large-scale NLI datasets, e.g. Marelli et al. (2014); Khot et al. (2018); Conneau et al. (2018), among others, typically use similar processes to create and label hypotheses from given premises. Here, we study models trained on MNLI, SNLI, plus other notable datasets including ANLI (Nie et al., 2020a) and VitaminC (Schuster et al., 2021). We aim to understand the behavior of models trained on these datasets at recognizing relevance between hypothesis and premise pairs.

## 3 A Case Study of Reference (In-)Determinacy

NLI models are typically finetuned exclusively on examples created with the reference determinacy assumption. We first study the effect of the RD assumption when we use such NLI models to solve downstream tasks. Specifically, we aim to understand how an NLI model would behave in a realistic scenario where the premise can be irrelevant to the hypothesis. In such cases, if there exists enough information in the evidence to establish reference determinacy, i.e. humans would be able to determine whether the evidence is related to the claim or not, an ideal NLI model should be able to correctly derive the NLI label.

Motivated by this, we study the use of NLI for the task of fact verification. We construct the RefNLI benchmark, which features 1,143 pairs of claim and retrieved Wikipedia evidence sentence, with human-labeled reference determinacy and entailment relations.

### 3.1 Sampling Claims and Evidence

We start by sampling claims from the validation and test splits of FEVER (Thorne et al., 2018) and VitaminC (Schuster et al., 2021). With each claim, we use BM25 to retrieve the top-10 passages from an English Wikipedia dump from 2018-07-01 with pyserini (Lin et al., 2021) [2]. Note that most of the retrieved passages would not be related to the entity or event described in the claim. Next, given each claim and each sentence in the top-10 retrieved passages, we classify their relation with a widely-used, pretrained RoBERTa model (Liu et al., 2019) finetuned on a mixture of NLI datasets from Nie et al. (2020a).

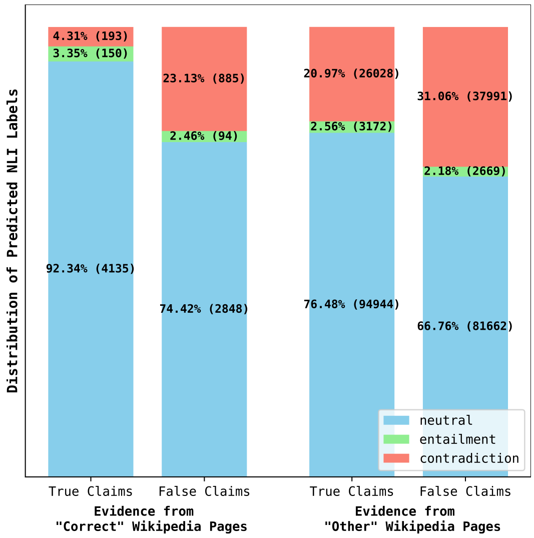

Figure 1: The distribution of label predictions by RoBERTa NLI mixture model from Nie et al. (2020a) when used to verify claims against retrieved evidence sentences from the correct vs. (most likely) irrelevant Wikipedia pages.

**NLI model predicts many false contradictions.** On the development set of FEVER, we compare how the model behaves when a claim is verified against evidence sentences from the "correct" Wikipedia page labeled in FEVER, compared to sentences from other Wikipedia pages, which are more likely to be irrelevant to the claim. Figure 1 shows the distribution of the NLI model's label predictions when used to verify claims labeled as *supported (True)* or *refuted (False)* by Wikipedia in FEVER. We observe that apart from the case where true claims are verified against sentences from the correct Wikipedia page, NLI models make contradiction predictions much more frequently than entailments in all the other three cases. While finding contradictions of false claim in the *correct* Wikipedia page where the refuting evidence comes from is what we want to see, interestingly we observe that the NLI model predict much more contradictions against *irrelevant* Wikipedia pages, i.e. pages about a different entity. In cases where

[2] https://github.com/castorini/pyserini

| Hypothesis | Premise | Label | NLI prediction |
|---|---|---|---|
| Sabbir Khan made his directorial debut in 2001. | In 2009 he made his directorial debut with the film "Kambakkht Ishq" (2009) that starred Akshay Kumar and Kareena Kapoor. | Ambiguous | Contradiction |
| Wales has a large region rich in coal deposits. | Recent explorations have revealed prospective deposits of rare-earth elements, a company is proposing further analysis of these mineral deposits. | Neutral | Contradiction |
| Same Old Love is a work of music. | "Same Old Love" was also performed on "The Ellen DeGeneres Show", "The Tonight Show Starring Jimmy Fallon", 2015 American Music Awards, and at the 2015 Billboard Women in Music. | Entailment | Contradiction |
| Buffy the Vampire Slayer is exclusively a Japanese television series. | "Buffy the Vampire Slayer" comics refer to comic books based on the television series "Buffy the Vampire Slayer" | Contradiction | Entailment |

Table 2: Examples from our study and the REFNLI benchmark. Compared to the usual three-way NLI label set, i.e. *entailment*, *neutral* and *contradiction*, we explicitly distinguish the *ambiguous* cases, where reference determinacy between the hypothesis and premise is meaningful yet cannot be established. "NLI predictions" shows predictions made by the RoBERTa-based NLI model Nie et al. (2020a) under three-way classification.

the sentence comes from such irrelevant Wikipedia pages, the pattern of potential "false contradictions" from the model is largely visible. The finding here echoes our initial hypothesis, suggesting the NLI model seems to be lacking the ability to recognize whether an evidence sentence refers to the same context as the claim.

### 3.2 REFNLI Benchmark

To further validate our hypothesis and understand why NLI models behave this way, we design a human study and analyze the example predictions made by NLI models in this setting.

From the set of examples where the RoBERTa NLI model predicts entailment or contradictions, and the evidence does not come from the correct Wikipedia page, we sample a subset for human annotation uniformly at random. The authors of the paper then annotate each claim and evidence sentence pair with one of these four labels:

- *Entailment*: if the human annotator thinks that the evidence and claim likely refer to the same context, and the evidence is sufficient to fully support the claim.
- *Contradiction*: if the human annotator thinks that the evidence and claim likely refer to the same context, and claim is unlikely to be true given the evidence.
- *Ambiguous*: if it is unclear whether the claim and the evidence refer to the same context (e.g. contain ambiguous reference), and there exist multiple possible assignments or interpretations of references that could make the example fall into at least 2 of the other 3 labels. *Neutral*: if it is clear that the evidence cannot support or contradict the claim in any way, i.e. there exists no interpretation or assignment of references of the evidence where it can support or contradict the claim.

Compared to the usual 3-way NLI labels, the label set here is designed to distinguish where reference determinacy cannot be safely established between a hypothesis and a premise. Note that even in such cases, the label could still be *neutral*, as long as the premise is unrelated to the hypothesis, no matter how the ambiguous reference is interpreted. This follows the intuition that ambiguity in reference determinacy only matters when there exists an interpretation where the evidence could be related to the claim. To help understand the motivation behind the label set design, we include one example of each label in Table 2. For instance, the first example is labeled as ambiguous, as it is not clear who *he* refers to in the premise, unless we assume the hypothesis and premise must be talking about the same entity. We include a more detailed description of the annotation guidelines and discussion of corner cases in Appendix A.

**The difference between *neutral* vs. *ambiguous*.** From the NLI task's perspective, the notable difference is that *neutral* hypothesis-premise pairs themselves contain enough information for humans to judge that the premise is irrelevant to the claim. In such cases, it is reasonable to expect a good NLI model to make the correct prediction, whereas for *ambiguous* examples, the correct label cannot be determined without the RD assumption. In our study, we do not expect NLI models to work well for ambiguous examples. NLI models' behavior with respect to ambiguity is investigated in greater detail in a recent study from Liu et al. (2023).

**Annotation process.** The authors went through a total of 1,143 example pairs, where one author produced the initial label and another author verified and adjudicated the label. On a sub-sample of 102 claims, we ask three authors to produce the label individually and we observe 0.83 Fleiss' $\kappa$ under 4-way classification, suggesting a good inter-rater agreement under the setting. In the rest of the paper, we denote the annotated set of examples as the REFNLI benchmark.

**Statistics.** In REFNLI, the authors went through a total of 1,143 pairs of claim and evidence sentences, with 905 *neutrals*, 66 *contradictions*, 37 *entailments*, and 135 *ambiguous* cases.

## 4 Evaluating Model's Reference Determinacy Biases

With REFNLI, we try to understand the effect of training datasets on the resulting NLI models' capabilities of recognizing reference determinacy. For this, we finetune a T5-large (Raffel et al., 2020) model on different combinations of NLI datasets, and study their behaviour on REFNLI.

### 4.1 Experimental Settings

**Datasets.** We study a mixture of five large-scale NLI datasets: SNLI (Bowman et al., 2015) MNLI, (Williams et al., 2018), ANLI, (Nie et al., 2020a) and VitaminC (Schuster et al., 2021) and the processed NLI sentence-pair style of FEVER used in VitaminC.

**Training.** We initialize the model with pretrained T5-large 1.1 checkpoint using the T5x library (Roberts et al., 2022). We finetune the model with different combinations of the datasets, as shown in Table 3. The label set across dataset is unified to match the three-way classification on MNLI and SNLI, where each label is represented as a single token in the T5 output vocabulary space. For variations of training dataset (mixtures), we use a learning rate of $1e-4$ with the Adam optimizer (Kingma and Ba, 2014) and batch size of 128 during finetuning.

**Evaluation.** We evaluate each finetuned model on all examples in REFNLI. We report the per-label precision and recall of predicted label, which is computed by the output label token with the highest softmax probability. To account for the effect of using different classification thresholds for each label in label imbalanced setting, we additionally report the per-label area under ROC (AUROC) score over the output label probability distribution under one-label-vs-rest setting.

We additionally evaluate Gemini$_{\texttt{ultra}}$ with 8-shot in context learning (GTeam et al., 2023) as a point of comparison to contrast the behavior of finetuned NLI models with an instruction tuned large langauge model.

### 4.2 Results

Table 3 shows the classification results. We generally observe that models exhibit low precision and high recall on both contradiction and entailment predictions, suggesting the presence of many false positive predictions made on the two labels. In terms of AUROC, it's more visibly clear that models perform generally worse on recognizing contradictions compared to recognizing entailments, which echoes our observations in §3.

**All training datasets show similar patterns of false contradictions and entailments.** Across all combinations of training datasets, we observe similar patterns of many false contradiction and entailment predictions, with slight variations across datasets. With respect to entailment predictions, we see almost all training configurations lead to high AUROC score (i.e. $> 0.85$). However, with respect to contradictions, we observe a larger discrepency across different datasets. We observe that including SNLI and Fever(NLI) in the training mix would lead to worst performance in terms of contradiction detection. In both leave-one-out and single dataset training settings, we observe ANLI to be the most useful dataset to include during training, especially for contradiction detection. Interestingly, ANLI (arguably) happens to be the one dataset where the reference determinacy assumption is least enforced during the annotation process, yet no definitive conclusion can ever be drawn here due to the existence of many other confounders.

On Gemini$_{\texttt{ultra}}$, we observe a much lower rate of false contradiction and entailment compared to all of the finetuned NLI models. That said, there still exists a gap between the performance on con-

| Training Data (Model) | Contradiction (66) | | | Neutral (905) | | | Entailment (37) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | AUROC | Precision | Recall | AUROC | Precision | Recall | AUROC |
| ALL | 15.76±3.74 | 92.42±6.73 | 90.91 | 98.99±0.92 | 53.92±3.36 | 87.49 | 25.78±7.92 | 89.18±10.75 | 94.53 |
| ALL - SNLI | 17.39±4.04 | 90.91±7.18 | 90.51 | 98.80±0.88 | 63.75±3.23 | 89.25 | 36.71±10.95 | 78.37±13.56 | 91.74 |
| ALL - MNLI | 22.30±5.23 | 90.91±7.15 | 89.62 | 98.93±0.80 | 71.93±3.09 | 89.42 | 38.27±10.43 | 83.78±12.19 | 94.68 |
| ALL - ANLI | 19.87±4.58 | 89.39±7.61 | 88.50 | 98.46±1.00 | 70.60±3.09 | 88.02 | 50.00±12.83 | 83.78±12.33 | 93.92 |
| ALL - Fever(NLI) | 15.71±3.58 | 90.81±7.14 | 87.85 | 98.35±1.11 | 59.44±3.28 | 84.47 | 37.97±11.23 | 81.08±13.32 | 96.13 |
| ALL - VitaminC | 14.06±3.39 | 92.42±6.76 | 88.37 | 98.85±10.35 | 47.40±3.31 | 83.68 | 23.57±6.99 | 89.19±10.73 | 94.57 |
| SNLI | 8.40±2.08 | 93.94±5.93 | 72.21 | 97.07±2.33 | 21.99±2.71 | 66.46 | 50.77±12.16 | 89.18±9.85 | 96.70 |
| MNLI | 10.91±3.69 | 93.94±8.04 | 88.48 | 98.20±0.93 | 42.21±2.32 | 81.17 | 62.75±9.57 | 86.48±7.88 | 94.93 |
| ANLI | 19.04±4.52 | 90.91±7.19 | 92.18 | 98.85±0.86 | 66.96±3.12 | 91.66 | 38.75±10.91 | 83.78±12.03 | 95.60 |
| Fever(NLI) | 6.29±4.08 | 13.64±8.90 | 58.04 | 90.42±2.46 | 66.74±8.90 | 57.67 | 12.69±5.21 | 67.57±14.63 | 85.10 |
| VitaminC | 19.64±3.80 | 83.33±9.59 | 87.49 | 98.23±1.05 | 67.40±3.06 | 85.76 | 28.97±8.21 | 83.78±12.12 | 93.19 |
| Gemini$_{\text{1.0 - Ultra}}$ | 36.79±4.23 | 59.09±6.35 | - | 96.46±1.13 | 90.49±2.74 | - | 56.60±10.15 | 81.08±11.42 | - |

Table 3: Per-label classification precision and recall on REFNLI from T5-Large finetuned on different combinations of five NLI datasets, and Gemini$_{\text{ultra}}$ with 8-shot prompting for comparison. ALL denotes using the mixture of all five datasets for finetuning, and ALL − X denotes the leave-X-out mixture. We generally observe that all combinations of training data leads many false contradiction and false entailment in predictions. Number in parathenses shows label count in the benchmark. ± shows 95% confidence interval of precision and recall, estimated via bootstrap resampling with 500 iterations. All metrics shown are scaled by 100× for visualization purposes.

| Model | $F_1$ score w.r.t each label | | |
|---|---|---|---|
| | *Entails* | *Neutral* | *Contradicts* |
| T5-Small | 84.14 | 84.64 | 78.02 |
| T5-Base | 88.91 | 88.42 | 82.36 |
| T5-3B | 93.79 | 92.19 | 87.95 |
| BERT-Tiny | 71.78 | 75.65 | 68.09 |
| BERT-Base | 85.85 | 85.88 | 80.10 |
| BERT-Large | 89.13 | 88.11 | 82.63 |

Table 4: Per-Label $F_1$ score of different models finetuned on MNLI and tested on MNLI validation set. We observe that model generally perform worse on contradictions compared to the other two labels.

| Label | Metric | | |
|---|---|---|---|
| | *Precision* ↑ | *Recall* ↑ | *AUROC* ↑ |
| Entail. | 15.76 → **32.26** | **89.18** → 84.85 | 90.91 → **94.57** |
| Neutral | **98.99** → 97.81 | 53.92 → **69.09** | 87.49 → **88.49** |
| Contra. | 15.76 → **20.29** | **92.42** → 84.85 | 90.91 → **91.18** |

Table 5: Per-Label precision recall and AUROC of T5-large trained on the mixture of five datasets before → after training set filtering described in §4.4

tradictions vs. entailments. For Gemini, we do not report the AUROC score as we do not have access to the output token probabilities during inference.

## 4.3 Are Contradictions More Difficult to Learn?

In the previous section, we observe a wide performance gap when finetuned NLI models are applied to recognize contradictions in settings where reference determinacy cannot be assumed. An additional factor here is that contradiction might be inherently a more difficult problem to learn from the training data distribution. Table 4 shows an experiment where we finetune different variants of BERT (Devlin et al., 2019) and T5 on the MNLI training set. When we evaluate the models on the MNLI dev set, we observe that the model consistently perform worse on contradiction examples. Here we hypothesize that the low validation performance of contradictions might be attributed to the inherent human disagreement (Pavlick and Kwiatkowski, 2019), where the human raters tend to have more disagreements on contradictions compared to the other labels. We show and discuss evidence of this, as well as how this can be connected to the reference determinacy assumption later in §5.

## 4.4 Mitigating the Effect of Reference Determinacy

To further validate that the reference determinacy assumption in the training data has an impact on downstream performance, we demonstrate that filtering out examples where reference determinacy cannot be easily determined improves the resulting model's performance on REFNLI.

With the mixture of five training datasets, we check whether a contradiction or entailment example is likely to be affected by the reference determinacy assumption, by the simple heuristics of lexical overlap. If a hypothesis and the premise share a token-level Jaccard similarity less than or equal to 0.15, we would discard this example from the training set, as we conjecture that it is more likely that the example is only labeled as contradiction or entailment due to the RD assumption. We filter out such examples from the training mix, and perform a rebalance of the label distribution by random re-

| Dataset | Ambiguous Reference? | Correlation Between Human Votes (↓) | | |
|---|---|---|---|---|
| | | *Ent.* ↔ *Neu.* | *Ent.* ↔ *Con.* | *Con.* ↔ *Neu.* |
| SNLI (Bowman et al., 2015) | *All* | -0.63$^{**}$ | -0.73$^{**}$ | **-0.08**$^{*}$ |
| | No (˜53%) | -0.74$^{**}$ | -0.48$^{**}$ | **-0.23**$^{**}$ |
| | Yes (˜47%) | **-0.36**$^{**}$ | -0.51$^{**}$ | -0.61$^{**}$ |
| MNLI (Williams et al., 2018) | *All* | -0.62$^{**}$ | -0.50$^{**}$ | **-0.37**$^{**}$ |
| | No (˜54%) | -0.64$^{**}$ | -0.74$^{**}$ | **-0.03** |
| | Yes (˜46%) | -0.52$^{**}$ | -0.70$^{**}$ | **-0.25**$^{**}$ |

Table 6: To understand how reference ambiguity affects human agreement in NLI, we compute the Pearson correlation among 100 human votes per example provided in ChaosNLI (Nie et al., 2020b). Correlation of $-1$ indicates perfect agreement among raters on the distinction between two labels, and vice versa. We randomly sample 500 examples respectively from SNLI and MNLI split of ChaosNLI and annotated whether each example contains ambiguous reference or not. (* denotes $p < 0.05$, ** denotes $p < 0.01$ for the correlation coeffecient.)

sampling neutral examples to match the number of contradiction or entailment examples left in the dataset.

The evaluation results are shown in Table 5. We see that the method generally improves the precision of entailment and contradiction predictions. We also see minor improvements across all labels in terms of AUROC. The findings here further validate our hypothesis that training with examples created with the RD assumption has a trickle-down effect on the performance of NLI models in real-world settings.

## 5 Can Reference (In-)determinacy Explain Human Disagreements?

Next, we study whether inherent human disagreements (Pavlick and Kwiatkowski, 2019) on NLI labels can potentially be attributed, at least in part, by the reference ambiguity between the hypothesis and premise. We conduct an experiment with the ChaosNLI dataset (Nie et al., 2020b). ChaosNLI contains samples of the original SNLI and MNLI datasets, where each example is re-labeled by 100 different crowdsource workers. ChaosNLI presents an interesting case for our purpose, as the human raters were not given explicit instructions to assume reference determinacy, which was instead deferred to their own judgement. To understand whether and how reference ambiguity might lead to human disagreements, the authors went through 500 random samples respectively from SNLI and MNLI split of ChaosNLI, and labeled whether ambiguity exists between the hypothesis and premise, following the same annotation protocol as in §3.2.

We compute the Pearson correlation between the number of votes each label received for each NLI example. Here, a higher correlation value between two labels (e.g., → 1) indicates that humans disagree and confound the two labels more often, and vice versa. Table 6 shows the our results.

**Humans disagree more between contradiction and neutral labels.** Overall, we observe that human raters tend to split votes between the neutral and contradiction labels more frequently than other combinations. Notably, on SNLI, we see a much weaker negative correlation ($r = -0.08$) between contradiction and neutral, compared to the relatively strong negative correlation between the other two label pairs. On MNLI, we observe a similar pattern, yet the gap is much smaller ($r = -0.37$ between contradiction and neutral). When we compare the ChaosNLI annotations against the original labels from MNLI and SNLI's five-way annotation, we observe that the change in majority label happens more often between entailment vs. neutral and contradiction vs. neutral, as shown in Figure 3 in Appendix B.

**Human disagreements can in part be attributed to reference ambiguity.** To estimate the percentage of examples that exhibit disagreements due to reference determinacy, we look at how the correlation between votes on different labels changes with respect to whether reference ambiguity exists in the data. From Table 6, we see that in both MNLI and SNLI, a large fraction of the examples exhibit the problem of reference ambiguity (˜47% in SNLI, ˜46% in SNLI). When we compare the case between ambiguous vs. unambiguous examples, we see that on both datasets, the rater agreement between contradiction and neutral improves when we go from ambiguous to unambiguous cases, while we observe the vice versa between entailment and neutral labels. We observe that the change in agree-

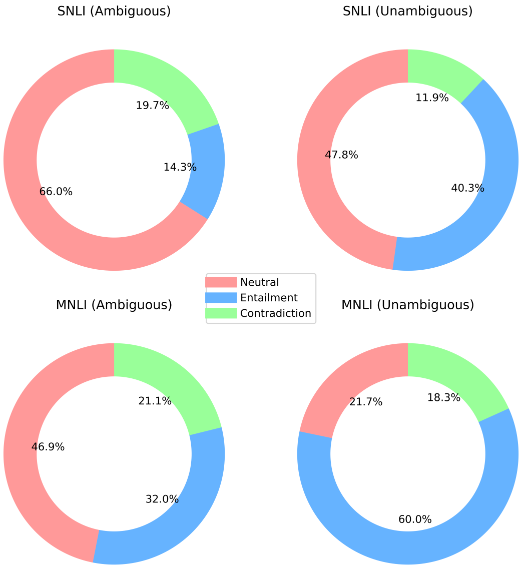

Figure 2: Distribution of the majority labels from the MNLI and SNLI split of ChaosNLI, when the reference between the hypothesis and premise is *ambigious* vs. *unambiguous*.

ment patterns are mostly due to whether the rater can safely establish reference determinacy between the hypothesis and premise. If so, then whether raters would agree on the hypothesis is contradicted by the premise is less likely to be impacted by the additional judgement of whether the two statements refer to the same context.

In Figure 2, we see how the majority label distribution shifts according to whether ambiguity exists in NLI examples. We observe that in ambiguous cases, the annotators are more likely to label an example as neutral, while in the unambiguous case, raters are more likely to judge the hypothesis as entailed or contradicted by the premise.

The findings here echo our hypothesis that the existence of reference ambiguity in NLI examples would lead to more disagreements among annotators. This potentially suggests that human disagreement can at least in part be attributed to the reference (in-)determinacy problem, and the annotation process would have more disagreement especially when raters are not explicitly instructed to assume RD during the annotation process.

## 6 Related Work

As ambiguity is an indispensable element in how we interpret and express language, many language understanding tasks require models to be able to recognize the resolve the ambiguity that exists in an user query (Xu et al., 2019; Zamani et al., 2020; Stelmakh et al., 2022; Feng et al., 2023; Zhao et al., 2024). For instance, Min et al. (2020) observe that ambiguous questions might lead to different answers depending on what the user intent is, and this would lead to annotation ambiguities when raters are asked to provide a single answer for an ambiguous question. With NLI, previous studies (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b) have found that inherent human disagreements exist in NLI labels, and the disagreement usually follows instance-dependent pattern. This work explores the understudied problem of explaining and understanding the cause of disagreements. Being able to understand the disagreements can potentially lead to the development of better NLI systems, as Zhou et al. (2022) and Zhang and de Marneffe (2021) show the merit of modeling the uncertainty distribution of NLI labels.

Our work tries to understand the impact of annotation artifacts (Gururangan et al., 2018; Bowman et al., 2020) on the downstream applicability of NLI tasks and models. In practice, researchers have found that NLI models would exploit such artifacts (Poliak et al., 2018; McCoy et al., 2019), which potentially hurts the downstream applicability. Our work is motivated by the use case of using NLI for verifying text and factual consistency (Schuster et al., 2021, 2022; Honovich et al., 2022; Gao et al., 2023), and we seek to understand the limitation of NLI models in such use cases.

Beyond NLI and its downstream applciations, it remains to be seen whether the reference or context ambiguity problem exists in other tasks and datasets as well. Along this line, Liu et al. (2023) designs a suite of tests that show current instruction-tuned language models often fail to respond to input ambiguity. We conjecture that this could be due to the inherent reference ambiguity in other tasks during the instruction-tuning stage of these models. We hope to explore this thread in future work.

## 7 Conclusion

This paper studies the impact of the reference determinacy assumption in the NLI dataset creation process. We release the RefNLI benchmark, and investigate the trickle-down effect of reference ambiguity in NLI on both the human annotators and subsequently on the NLI model training process. We hope that future NLI researchers and practitioners pay attention to this problem, especially when trying to apply NLI models in downstream use cases.

## Limitations

Our study focuses on understanding the implication of reference (in-)determinacy and its impact from a data perspective. Our modeling experiments use one fixed architecture with different mixtures of NLI datasets for training. Although it is mostly due to the fact that we want to understand the impact of using different types of NLI datasets for training, experimenting with more models could potentially eliminate model architecture as the confounder in our results. Although not the focus of our study, but the study could be extended and strengthened with experiments with large language models to understand the models react and respond to ambiguities in the input with the NLI task format. As we discussed at the end of §6, we leave the two parts for future exploration.

## Ethical Considerations

To the best of our knowledge, our study does not introduce ethical concerns.

## 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.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214, Online. Association for Computational Linguistics.

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

I. Dagan and O. Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

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.

Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. 2023. Generic temporal reasoning with differential analysis and explanation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12013–12029, Toronto, Canada. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

GTeam, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *AAAI Conference on Artificial Intelligence*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. 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.

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

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Alan Ritter, Stephen Soderland, Doug Downey, and Oren Etzioni. 2008. It's a contradiction – no, it's not: A case study using functional relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Honolulu, Hawaii. Association for Computational Linguistics.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020. Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3189–3196.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Xinran Zhao, Tong Chen, Sihao Chen, Hongming Zhang, and Tongshuang Wu. 2024. Beyond relevance: Evaluate and improve retrievers on perspective awareness. *arXiv preprint arXiv:2405.02714*.

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.

## A RefNLI Annotation Guidelines

The expert raters for RefNLI were presented with examples consisting of a premise and a hypothesis. For each example, they were given instructions as follows.

You are to assign one of 4 labels to the example:

(a) **ambiguous reference**: If the premise contains ambiguous reference, and it's possible that with resolved reference, premise would actually support/contradict the claim.

(n) **neutral**: If the premise can't support or contradict the claim in any possible way. e.g. No matter how you resolve the reference, the premise would still be irrelevant to the claim.

(c) **contradiction**: If the claim is most likely false given the premise.

(e) **entailment**: If premise fully supports the claim.

If you find tricky cases, put yourself in the following scenario: Suppose an LLM generates the claim, you want to decide if we should, given the evidence, tell the user that that this claim is true, tell the user that it's false, or neither.

The distinction between neutral and ambiguous is going to be difficult sometimes. See examples below for what we are after. If it's truly unclear – feel free to skip the example.

### Specific Guidelines

1. **Skip unclear claims or premises**: If you think the claim is difficult to understand, or there is too much ambiguity, skip the claim entirely.

2. **Don't label the claim by its truth value in the world**: If a claim says "The sky is blue", and the premise says something completely different, label it as neutral. Don't label such cases as entailment based on **just** your world knowledge.

3. **World Knowledge is permitted**: You can assume commonly accepted world knowledge when interpreting the premise, e.g., basic geography and other commonsense knowledge are allowed. If needed, a web search is allowed when making the judgements. However, don't make too many inferences.

4. **Temporal considerations**: Ignore tense (e.g., past or present) in both the premise and claims. If the premise clearly indicates a time of an event, but the claim doesn't, assume that the claim is uttered right after the event.

5. **Personal surnames**: If only the surname of a person is mentioned in the premise, and there's not enough evidence for in the premise for you to determine the last name is referring to the same entity as in the hypothesis, mark the example as "ambiguous"

6. **Neutral vs. Ambiguous Reference**: The distinction between the two can be difficult sometimes. The general rule is: if the premise can't seem to support the claim no matter how you interpret the premise, then it's neutral.

### Some examples given in the instructions

**Premise**: Wales has a large region rich in coal deposits.
**Hypothesis**: The Ural Mountains contain about 48 species of economically valuable ores and economically valuable minerals.
**Label**: N; even if we didn't know whether the Ural Mountains are in Wales, the premise doesn't mention anything about coal deposits, so there's no way that the premise can support/contradict the claim.

**Premise**: Wales has a large region rich in coal deposits.
**Hypothesis**: Famous for its coal, Newcastle is the largest coal exporting harbour in the world, exporting 159.9 million tonnes of coal in 2017.
**Label**: A: The prominent Newcastle is in New South Wales, Australia, but there happens to also be a small town named Newcastle in Wales

**Premise**: The Predator made more than $97 million worldwide.
**Hypothesis**: Up to March 2011, The Predator's worldwide gross has reached $172,543,519, making it the highest-grossing film in the franchise.
**Label**: E; if the premise mentions a time, and there's no clear temporal marker in the claim – assume that the claim is made in the similar time frame as the premise.

**Premise**: The Hunchback of Notre Dame is a Disney media franchise, commencing in 1996 with the release of "The Hunchback of Notre Dame".
**Hypothesis**: The Hunchback of Notre Dame has only ever been based off of a poem.

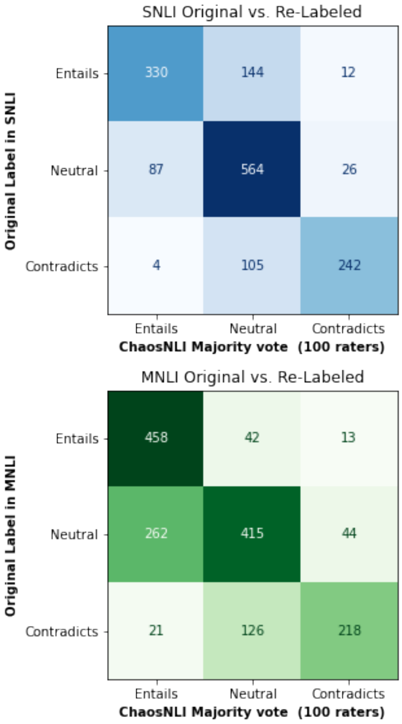

Figure 3: Confusion Matrices between majority label from the original annotation vs. ChaosNLI's re-annotation label for SNLI and MNLI examples from Nie et al. (2020b).

**Label**: Skip, since it's unclear in the hypothesis what "based off of a poem" means

## B Human Disagreements and Reference Ambiguity

Figure 3 shows the confusion matrix between the majority NLI label from the ChaosNLI re-annotation vs. the original majority label from the five SNLI/MNLI annotators originally.