Thinking Like a Skeptic: Defeasible Inference in Natural Language

Rachel Rudinger[†‡§], **Vered Shwartz**[†‡], **Jena D. Hwang**[†], **Chandra Bhagavatula**[†], **Maxwell Forbes**[†‡], **Ronan Le Bras**[†], **Noah A. Smith**[†‡], **Yejin Choi**[†‡]
[†]Allen Institute for AI
[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[§]University of Maryland, College Park, MD
rudinger@umd.edu
{vereds,jenah,chandrab,maxf,ronanlb,noah,yejinc}@allenai.org

Abstract

Defeasible inference is a mode of reasoning in which an inference (X is a bird, therefore X flies) may be weakened or overturned in light of new evidence (X is a penguin). Though long recognized in classical AI and philosophy, defeasible inference has not been extensively studied in the context of contemporary data-driven research on natural language inference and commonsense reasoning. We introduce Defeasible NLI (abbreviated δ-NLI), a dataset for defeasible inference in natural language. δ-NLI contains extensions to three existing inference datasets covering diverse modes of reasoning: common sense, natural language inference, and social norms. From δ-NLI, we develop both a classification and generation task for defeasible inference, and demonstrate that the generation task is much more challenging. Despite lagging human performance, however, generative models trained on this data are capable of writing sentences that weaken or strengthen a specified inference up to 68% of the time.

1 Introduction

Commonsense reasoning tasks are frequently formulated in terms of *soft* inferences: what is *likely* or *plausibly* true given some context, rather than (or in addition to) what is *necessarily* true. Given a context such as "*The drinking glass fell*", it is common sense to infer that what likely happened next is "*The drinking glass broke*". However, with the addition of new information, this inference may be blocked or weakened. If, for example, we subsequently learn that "*The glass fell onto a pile of laundry*" or that "*The glass was made of durable material*", our original expectation that the glass will break is greatly diminished. This pattern of reasoning, in which an initially supported inference may subsequently be weakened or retracted in light of new evidence, is known as **defeasible reasoning** (Koons, 2017).

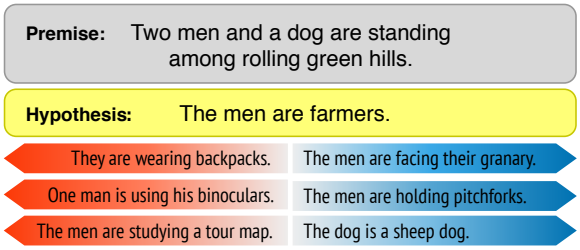

Figure 1: Examples from the δ-SNLI portion of the δ-NLI dataset. A neutral premise-hypothesis pair from SNLI is augmented with three update sentences that weaken the hypothesis (left, red) and three update sentences that strengthen it (right, blue).

To the extent, then, that commonsense and natural language inference systems must be able to reason about plausible or likely inferences, they must also be able to reason about the *defeasibility* of those inferences. While most contemporary resources and datasets for these tasks attempt to directly address the former, few provide the context to facilitate the latter mode of reasoning.

Tasks like the Recognizing Textual Entailment (RTE) challenge (Dagan et al., 2005) or Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) capture entailment relations between a fixed context (premise) and inference (hypothesis), but do not reveal how these relations may shift in light of new information about the context. Similarly, knowledge graphs for commonsense reasoning like ATOMIC (Sap et al., 2019) or ConceptNet (Havasi et al., 2007; Speer et al., 2017) encode inference rules about generic situations, but do not elaborate on possible exceptions to the applications of those rules.

In this work, we introduce Defeasible NLI (abbreviated δ-NLI, pronounced "delta-NLI"), a new dataset to study defeasible inference in natural lan-

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4661–4675

Task	Premise	Hypothesis	Type	Update
δ-SNLI	Old man crafting something in his workshop.	An old man is working .	strengthener	The man is serious and is surrounded by workers.
			weakener	The man is wearing pajamas and is chuckling.
δ-ATOMIC	PersonX has a pool party	Because PersonX wanted to hangout with friends	strengthener	It was PersonX's birthday
			weakener	PersonX was having a family reunion
δ-SOCIAL		You should help your family with funeral expenses.	strengthener	They have asked you to chip in
			weakener	You are not financially stable to help out

Table 1: Examples of strengtheners and weakeners collected for the δ-SNLI, δ-ATOMIC, and δ-SOCIAL portions of the Defeasible NLI dataset.

guage.[1] δ-NLI is a collection of extensions to three existing English-language inference datasets, covering a broad range phenomena: natural language inference (SNLI (Bowman et al., 2015)), commonsense reasoning (ATOMIC (Sap et al., 2019)), and reasoning about social norms (SOCIAL-CHEM-101 (Forbes et al., 2020)). We refer to these subsections of the dataset as δ-SNLI, δ-ATOMIC, and δ-SOCIAL, respectively. We augment each resource by eliciting additional contextual information ("updates") that either strengthen or weaken a given inference (which we term "strengtheners" and "weakeners," respectively). An example is provided in Fig. 1.

From these three augmented datasets, we are able to devise two tasks for defeasible inference: (1) a *classification* task for predicting whether a provided update sentence acts as a strengthener or a weakener; and (2) a *generation* task in which a premise-hypothesis pair are provided as input and an update sentence that weakens or strengthens the hypothesis must be generated as output. Through experiments in which we fine-tune pretrained language models for both tasks, we demonstrate that the generative task is much more challenging than the classification task. While system performance approaches human-level agreement on the classification task, the gap between system and human performance on the generative task is still considerable. We perform an extensive analysis of the failure and success modes of the generative defeasible inference models.

Finally, we observe that, not only is the generative task more challenging than the classification task, but it has an additional meaningful interpretation, namely, a system's ability to "think like a skeptic." That is to say, informally, a human who is engaging in skeptical reasoning is considering the possible weaknesses of a given claim or argument in order to come up with examples or counterarguments that may undermine it; by analogy, the generative task we introduce here requires a model to come up with (rather than simply verify) examples of circumstances that undermine the given hypothesis.

2 Background and Related Work

Defeasible reasoning is soft inference based on default assumptions to account for unknown facts, for example, "Tweety is a bird" entails that "Tweety flies", because birds usually fly. Such a conclusion is not deductively valid, and might be invalidated by new information such as "Tweety is a penguin" (Reiter, 1980; Lascarides and Asher, 1991). Defeasible reasoning is a type of nonmonotonic logic, as it contrasts the monotonicity property of classical logic, according to which valid inferences cannot be defeated by adding additional information (Kraus et al., 1990). Defeasible reasoning has been studied in a range of fields from logic, through linguistics and artificial intelligence.

Classical AI. In early AI, defeasible reasoning was used as a solution to the "frame problem": it is impossible to list all the potential effects of actions without describing mundane and obvious effects (McCarthy and Hayes, 1969). McDermott and Doyle (1980) offered a formal account of the proof systems and model theories of nonmonotonic

[1] Data available at https://github.com/rudinger/defeasible-nli

logics. Default logic (Reiter, 1980) was suggested as a nonmonotonic logic that specifies a set of default assumptions, i.e., predicates that are true unless specified otherwise (e.g., $\text{bird}(X) \rightarrow \text{fly}(X)$). In circumscription (McCarthy, 1980), defaults are expressed in natural language ("a bird will fly if it is not abnormal"). Pollock (1987) outlined a system for defeasible reasoning based on several different types of *warranted* inferences. Finally, Levesque (1990) suggested a special 'all I know is ...' operator, e.g. "All I know is that Tweety is a bird" entails that "Tweety flies".

Linguistics. In semantics and pragmatics, a distinction is drawn between *entailments* and *implicatures*. Entailments are inferences which are necessarily true, arising from the semantics of an utterance (e.g., "Pat is a bachelor," entails "Pat is unmarried."). Linguistic utterances also invite unstated pragmatic inferences, or *implicatures*, which depend not only on the semantics of the utterance but also its conversational context (Grice, 1975). Implicatures are *cancellable* (defeasible), meaning they could be revoked in light of further evidence. For instance, the comment "that cake looks delicious" might invite the inference that the speaker is requesting a slice, until they clarify that they have a food allergy. Building on this notion of default assumptions, Lascarides and Asher (1993) proposed to interpret discourse relations by defining defeasible rules based on commonsense knowledge of typical causes and effects.

Natural Language Processing. Textual entailment was defined as a softer version of semantic entailment, doubly hedging it with "a human would *typically think* that the hypothesis is *likely* true" (see Section 3, Dagan et al., 2005). It gained tremendous popularity again 10 years later, with the release of the large-scale Stanford Natural Language Inference dataset (SNLI; Bowman et al., 2015), that facilitated training neural models, and which was followed by several other datasets in that nature (Williams et al., 2018; Nie et al., 2019). But—among other criticisms of the task—it has been shown that people generally don't agree on entailment annotations (Pavlick and Kwiatkowski, 2019), and new variants of the task suggested to shift away from categorical labels to ordinal or numeric values denoting plausibility (Zhang et al., 2017; Sakaguchi and Van Durme, 2018; Chen et al., 2020). In this paper we focus on the *defeasibility* of textual entailments, a less well-studied phenomenon in this context.

3 Definition

In this paper, we employ a working definition of defeasible inference that may be seen as an outgrowth of prior work. Dagan et al. (2005) introduced the following informal definition for the Recognizing Textual Entailment (RTE) task:

> ...*textual entailment* is defined as a directional relationship between pairs of text expressions, denoted by *T*, the entailing "Text", and *H*, the entailed "Hypothesis". We say that *T* entails *H* if, typically, a human reading T would infer that H is most likely true.

Similarly, the task of Natural Language Inference (NLI) seeks to determine whether a (one-directional) entailment relation exists between a *premise* sentence and a *hypothesis* sentence (MacCartney, 2009; Bowman et al., 2015).

While the RTE and NLI tasks treat entailment relations as fixed, in this work we seek to understand how the introduction of new information can dynamically and directionally affect the strength of inference. Thus, our working definition of defeasible inference extends the RTE and NLI task formulations to model the relationship between a *premise*, *hypothesis*, and a third *update* sentence:

> Given premise $\mathcal{P}$, a hypothesis $\mathcal{H}$ is **defeasible** if there exists an update U (consistent with $\mathcal{P}$) such that a human would find $\mathcal{H}$ less likely to be true after learning $\mathcal{U}$. Specifically, an update $\mathcal{U}$ is called a **weakener** if, given a premise $\mathcal{P}$ and hypothesis $\mathcal{H}$, a human would most likely find $\mathcal{H}$ *less likely to be true* after learning $\mathcal{U}$; if they would find $\mathcal{H}$ *more likely to be true*, then we call $\mathcal{U}$ a **strengthener**.

By introducing *both* strengtheners and weakeners, we generalize from defeasibility as a one-directional phenomenon (weakening only) to study the bi-directional phenomenon.

4 Data Sources

We collect strengtheners and weakeners for three different types of data sources that illustrate the generality of the defeasible inference framework. Table 1 shows example strengtheners and weakeners collected for the various tasks, detailed below.

Natural Language Inference

The SNLI dataset (Bowman et al., 2015) is a large-scale human-labeled dataset created for natural language inference. It is a collection of 570K crowd-sourced English premise-hypothesis sentence pairs, each hypothesis manually classified as *entailment*, *contradiction*, or *neutral* with respect to its premise. The *neutral* pairs are of central interest in this work. In SNLI, neutral premise-hypothesis pairs are those in which the hypothesis is neither entailed nor contradicted by the premise (see Figure 1 for example), leaving room for the potential for strengthening or weakening the statement if the appropriate conditions are provided. In our dataset we include 10K *neutral* premise and hypothesis pairs, as well as a small subset of instances that lacked annotation consensus.[2]

Commonsense Knowledge Graph

The ATOMIC knowledge graph is a collection of 877K textual commonsense descriptions for inferential knowledge (Sap et al., 2019). The data was collected through crowdsourcing *if-then* knowledge about events and their commonsense relations to other events and states (relation targets). In ATOMIC, an event involving a PersonX is linked to multiple relation targets via relation types like `xAttr` (attribute of PersonX). For example, *if "PersonX adopts a cat"*, *then* PersonX might take a subsequent action (`xEffect`; *"buy cat litter"*), be seen as of a particular persona (`xAttr`; *"as seeking companionship"*), or have a particular mental state as a result (`xReact`; *"feels loved"*). While these relations capture commonsense inferences that are plausible or even very likely, their likelihood could be dampened with additional context, e.g., in the above case *"PersonX needs a barn cat for their mice problem"*. Thus, for the purposes of this study, we cast events as the premise and the relation targets as the defeasible hypotheses. In particular, we extract a total of 24K event (premise) and relation target (hypothesis) pairs. We limit the relation targets to six of nine relations corresponding to the explicit agent or PersonX in the event. The other three relations that concern 'others', which may or may not be explicit participants in the event, are excluded.

[2]Instances that were labeled '-' in SNLI.

Task	Split	#$\mathcal{P}$-$\mathcal{H}$	#S	#W
δ-SNLI	train	9,588	44,621	44,055
	dev	195	903	882
	test	203	924	913
δ-ATOMIC	train	19,518	17,662	17,340
	dev	2,155	1,937	1,903
	test	2,327	2,091	2,047
δ-SOCIAL	train	7,893	39,675	37,341
	dev	979	4,822	4,521
	test	982	4,867	4,572

Table 2: Number of unique $\mathcal{P}$-$\mathcal{H}$ pairs, strengtheners (S) and weakeners (W) in each section of the δ-NLI dataset.

Elicited (Human)		Predicted (Human Validation): Weakener	Strengthener	Neutral	None
δ-SNLI	Weakener	77.2	8.9	1.0	12.9
	Strengthener	3.0	89.7	2.0	5.4
δ-ATOMIC	Weakener	69.0	11.1	1.9	18.1
	Strengthener	2.6	87.4	0.4	9.6
δ-SOCIAL	Weakener	84.6	3.6	1.6	10.2
	Strengthener	1.8	90.1	1.4	6.2

Figure 2: Confusion matrix of human validation. Rows: the original update type for which updates were elicited. Columns: the update type that workers categorized them into during the validation step. Cells: percent of assignment into each category. "None" indicates no agreement between the annotators.

Statements of Social Norms

The SOCIAL-CHEM-101 dataset of social norms (henceforth, Social Norms) compiles a collection of 292K crowdsourced natural language statements about commonsense social judgments made given everyday situations (Forbes et al., 2020). These statements represent generic commonsense hypotheses about social behaviors and their acceptability that are held as norms in a society. However, such normative judgments can also be strengthened or weakened given appropriate context. For example, a norm like "It is good to respect your parents" might be weakened in certain contexts (e.g., "Your parents are abusive and hurtful towards you") and strengthened in others (e.g., "Your parents want what's right for you"). In other words, we consider this set of norms of social behavior as hypotheses capable of being strengthened or weakened. For our dataset, we randomly extract 10K statements of social norms.

5 Data Collection

Our data collection is performed via crowdsourcing (§5.1) and consists of two steps: update sentence elicitation (§5.2) and validation (§5.3).

5.1 Crowdsourcing

We carry out both the elicitation and validation steps via crowdsourcing in Amazon Mechanical Turk. To ensure the quality of our annotations, we have workers take a paid qualification test to assess their ability to follow instructions and to produce reasonable strengtheners and weakeners. The qualification test contains 6 manually selected premise-hypothesis pairs from SNLI that range from easy to difficult hypotheses to defeat. We then manually evaluate their responses for quality and adherence to the guidelines.

The 230 workers that provided acceptable updates (both strengtheners and weakeners) to a minimum of four test questions were selected to participate in the data collection tasks. Based on the feedback received from our worker pool, we updated the instructions with further clarifications and examples as necessary. Workers were paid over $15 per hour on average for all annotation tasks.

5.2 Update Sentence Elicitation

To collect update sentences for data sourced from SNLI and ATOMIC, we provide workers with a premise-hypothesis pair as prompt for which they are required to generate two free-text sentences: a strengthener and a weakener that will increase or decrease, respectively, the likelihood of the hypothesis being true. For the collection of updates for the Social Norms data, the workers are given the hypothesis and asked to provide two free-text sentences: a strengthener that supports the socio-normative assumption made in the hypothesis ("especially if...") and a weakener that undermines such assumption ("unless..."). Each elicitation HIT is performed by five workers.

In both cases, we provide the workers with the option to specify that a hypothesis cannot be updated. In order to prevent workers from creating incorrect or trivial updates, we require that the update does not contradict the premise, repeat or rephrase any of the premise or hypothesis, or simply negate the hypothesis.[3] We also instruct workers to avoid writing sentences that involve making stereotyped or prejudicial assumptions about people based on their identities (see §8 for additional information).

[3] See supplementary material for the complete HIT template.

Task	Inputs	RoBERTa	Human	Maj.
δ-SNLI	$(\mathcal{P}, \mathcal{H}, \mathcal{U})$	81.6	83.6	50.3
	$(\varnothing, \mathcal{H}, \mathcal{U})$	79.7		
	$(\varnothing, \varnothing, \mathcal{U})$	65.1		
δ-ATOMIC	$(\mathcal{P}, \mathcal{H}, \mathcal{U})$	78.3	78.2	50.5
	$(\varnothing, \mathcal{H}, \mathcal{U})$	77.7		
	$(\varnothing, \varnothing, \mathcal{U})$	65.2		
δ-SOCIAL	$(\varnothing, \mathcal{H}, \mathcal{U})$	86.2	87.6	51.6
	$(\varnothing, \varnothing, \mathcal{U})$	71.6		

Table 3: Accuracy (%) on the test set of each classification task.

5.3 Validation

In order to evaluate the validity of human annotations, we ask crowd workers to rate the collected strengtheners and weakeners with respect to the original premise-hypothesis pairs. The rating is on a 5-point Likert scale ranging from "weakens a lot" to "strengthens a lot" with a middle response category of "neutral" for those updates that have no update effect. Each validation HIT is annotated by three workers. The annotations yielded inter-annotator agreement with Krippendorff's $\alpha = 0.62, 0.67, 0.69$ for SNLI, ATOMIC and Social Norms, respectively (Krippendorff, 1980).

Figure 2 shows the results of the validation step. As evident, workers in the validation step successfully identified the intended update type of elicited updates, indicating the high quality of the elicited updates. In general, strengtheners showed higher agreement than weakeners.

The size of each dataset is given in Table 2. We assign instances into train, development, and test sets based on their split in the original datasets.

6 Defeasible Inference Tasks

We formulate two tasks: a *discriminative* defeasible inference task (§6.1) and a *generative* defeasible inference task (§6.2).

6.1 Defeasible Inference as Classification

We pose a binary classification task for defeasible inference: given a hypothesis $\mathcal{H}$, an optional premise $\mathcal{P}$, and an update $\mathcal{U}$, the goal is to determine the update type, i.e., whether $\mathcal{U}$ weakens or strengthens $\mathcal{H}$. That is, given an input tuple $(\mathcal{P}, \mathcal{H}, \mathcal{U})$, output a label in the set $\{\text{STRENGTHENER}, \text{WEAKENER}\}$.

To establish baseline performance, we fine-tune the transformer pretrained language model RoBERTa-base (Liu et al., 2019), which performs well in classification tasks, with a standard cross entropy loss, using the Transformers library (Wolf et al., 2019). We concatenate the sentences $\mathcal{P}$, $\mathcal{H}$, and $\mathcal{U}$ (separated by a special token) as input to RoBERTa, and select the best training run over five trials, run for two epochs each. Further training details are provided in the appendix. Following the hypothesis-only baseline suggested by Poliak et al. (2018), we also report the performance of versions of the model with partial inputs, i.e., $(\varnothing, \mathcal{H}, \mathcal{U})$ or $(\varnothing, \varnothing, \mathcal{U})$.

Results. Table 3 displays the classification accuracy on each task. For the models which have access to the full input $(\mathcal{P}, \mathcal{H}, \mathcal{U})$, accuracy is very close to human performance on each dataset. This suggests that *discriminating* between strengtheners and weakeners is a comparatively easy task for a strong pretrained language model like RoBERTa. For this reason, we primarily focus on the much more challenging task of *generating* strengtheners and weakeners, as described in the following subsection.

A partial explanation for the easiness of the classification task is due to annotation artifacts (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018), which are a known limitation of some text datasets directly elicited from human annotators. (See §8 for a full discussion of these limitations.) To check, we train a baseline with access to only the update as input. While this baseline performs 15 to 20 points above the uninformed majority baselines (indicating the presence of annotation artifacts), it is still 13 to 15 points below the fully-informed models.

Interestingly, removing *only* the premise (but not hypothesis) from the input only slightly decreases overall accuracy. This suggests most of the necessary signal is present in the hypothesis and update. See §7 for further discussion.

6.2 Generative Defeasible Inference

In the generative defeasible task, given a hypothesis $\mathcal{H}$, an optional premise $\mathcal{P}$, and a required update type (weakener or strengthener), the goal is to generate an update $\mathcal{U}$ that satisfies this constraint, i.e., weakens or strengthens $\mathcal{H}$.

We report the performance of several strong baselines, namely fine-tuning pre-trained transformer-based language models (LM). Specifically, we study two types of models: (1) encoder-decoder models such as Bart and T5 (Lewis et al., 2019; Raffel et al., 2019), which initialize the state of the transformer by reading the input, and learn to generate the output; and (2) standard LMs such as GPT and GPT2 (Radford et al., 2018, 2019), which are trained with the LM objective to predict the next word. We use the Transformers package (Wolf et al., 2019) and train each model for a single epoch with a batch size of 64. Further training details are provided in the appendix.

Task	LM	Ppl.	BLEU	ROUGE	Human Eval.
δ-SNLI	T5-large	2.51	11.48	25.03	
	Bart-large	2.46	**17.03**	**27.91**	38.22
	GPT2-S	5.18	12.61	26.35	
	GPT2-XL	3.84	15.66	27.78	**64.44**
	GPT2-XL $\mathcal{H}$-only	4.81	14.82	27.19	53.51
	Human				83.46
δ-ATOMIC	T5-large	4.58	0.89	12.13	
	Bart-large	2.04	4.13	20.40	30.21
	GPT2-S	3.21	3.74	18.14	
	GPT2-XL	2.20	4.77	**21.89**	**61.89**
	GPT2-XL $\mathcal{H}$-only	3.23	**5.59**	21.65	60.25
	Human				78.31
δ-SOCIAL	T5-large	2.69	4.33	14.94	
	Bart-large	2.09	10.65	18.43	34.49
	GPT2-S	2.53	9.56	16.97	
	GPT2-XL	2.11	**12.16**	**18.77**	**68.15**
	Human				86.60

Table 4: Automatic and human evaluation results on the test set, for the generative models. The input for each task was `[premise] p [hypo] h [type]` (`[hypo] h [type]` for Social Norms and for the hypothesis-only models), where `[type]` $\in$ {`[strengthener]`, `[weakener]`}, and **p** and **h** are the premise and hypothesis tokens, respectively. Note that the perplexities of the encoder-decoder models are not comparable to those of the standard LM models, since the latter compute the loss over the entire sequence, while the former compute the loss only over the output sequence. All models score well below human performance.

Automatic Evaluation. We follow the common practice of reporting automated generation evaluation metrics. We report the perplexity on the test set, as is often used to measure the performance of a language model.[4] In addition, we generated predictions for the test set using beam search with 5 beams, and evaluated them using standard n-gram based metrics: the precision-oriented BLEU-4 score (Papineni et al., 2002), which considers n-grams up to $n = 4$, and the recall-oriented

[4]Micro and macro perplexities were identical.

	Premise	Hypothesis	Type	Generated Update
①	A man just roaming on the streets during night.	A man is roaming the streets at night, drunk.	S	The man has a beer in his hand
	—	It is rude to point out their weight problem.	W	You are a nutritionist
②	PersonX pays PersonX's debt	Because PersonX wanted to be debt free	W	PersonX is in debt to the IRS
	—	It is rude to refuse help.	S	You are in an emergency
③	—	It is wrong to kill an animal.	S	You are trying to save the life of a child.
④	Old man crafting something in his workshop.	An old man is working at his home workshop.	W	The man is in his garage.
⑤	PersonX is on a fishing trip	PersonX then takes time.	W	PersonX is on a fishing trip.
⑥	Child squirting paint onto a piece of spin-art.	The child has paint all over their hands.	S	The child's hands are covered in paint.
⑦	PersonX picks PersonY up at the airport	Before, PersonX needed to get ready	W	PersonX is driving PersonY to the airport
⑧	PersonX looks at one another	PersonX is seen as threatening	S	PersonX is holding a gun
⑨	—	You should be respectful of people who just had a baby.	S	The baby is your first child.

Table 5: Examples of generations with update type (W = weakener, S = strengthener), across tasks and models, that fall into each of the nine analysis categories ① – ⑨ described in §7.

Task	Model	①	②	③	④	⑤	⑥	⑦	⑧	⑨
δ-SNLI	GPT2-XL	54.2	25.0	0.00	0.00	0.00	0.00	16.7	0.00	4.20
	Bart-large	25.0	29.2	4.17	12.5	0.00	29.2	0.00	0.00	0.00
	Overall	39.6	27.1	2.08	6.30	0.00	14.6	8.33	2.08	0.00
δ-ATOMIC	GPT2-XL	36.0	16.0	4.00	4.00	4.00	4.00	20.0	8.00	4.00
	Bart-large	24.0	20.0	8.00	12.0	20.0	4.00	4.00	4.00	4.00
	Overall	28.6	18.4	6.12	8.16	12.2	4.00	12.2	6.12	4.08
δ-SOCIAL	GPT2-XL	56.0	8.00	8.00	8.00	4.00	0.00	4.00	0.00	12.0
	Bart-large	32.0	24.0	8.00	24.0	0.00	4.00	0.00	0.00	8.00
	Overall	44.0	16.0	8.00	16.0	2.00	2.00	2.00	4.00	6.00

Table 6: Percentage distribution of generated updates over the analysis categories ① – ⑨ (described in §7), for each combination of task and model.

ROUGE-L score (Lin, 2004), which considers the longest common subsequences.

Table 4 presents the automatic evaluation results. We observe that the model preferences are consistent among BLEU and ROUGE. The GPT2-XL models perform best for δ-ATOMIC and the social norms dataset, and only slightly worse than the best model (Bart-large) on δ-SNLI. The model size does not have a major impact on performance, with GPT2-S performing moderately worse than GPT2-XL. The T5 model had the lowest performance across tasks in terms of BLEU and ROUGE.

Human Evaluation. Automatic metrics penalize models for lexical variability and often do not correlate with human judgements (Novikova et al., 2017). Thus, our main evaluation is human evaluation. The goal of the human evaluation is to determine the effectiveness of the models at generating weakeners and strengtheners, focusing on the best model in each category, namely GPT2-XL and Bart-large. We used the same crowdsourcing setup as the validation step in §5.3, and asked workers to rate the generated strengtheners and weakeners on a 5-point Likert scale.

Table 4 shows the human evaluation for Bart-large and GPT2-XL, in terms of accuracy score (e.g. a generated weakener was considered "correct" if the workers judged it as a weakener). As opposed to the automatic evaluation, in which these two models were comparable, here the outputs from GPT2-XL were judged as substantially better than Bart, but even so leaving room for improvement. Across models, strengtheners were not only easily agreed upon (§ 5.3) but also easier to predict than weakeners. In addition, the gap between the accuracy on strengtheners versus weakeners was narrower for GPT2-XL (17%) than for Bart (34%).

When applicable, we also report the performance of a hypothesis-only variant of the best-performing model (GPT2-XL $\mathcal{H}$-only in Table 4), for which the input consists of the hypothesis and the update type, excluding the premise. While this baseline performs similarly to the full model in terms of automatic metrics, the human evaluation reveals that the $\mathcal{H}$-only δ-SNLI model substantially underperforms the full model, suggesting that the generative model is making productive use of the premise in δ-SNLI; in the case of δ-ATOMIC, the disparity between the $\mathcal{H}$-only and full models is much smaller.

7 Analysis of Generated Updates

In order to analyze the quality of the generated updates, we sampled 150 instances from the development set (25 for each combination of task and model), and categorized their top prediction into the following categories, exemplified in Table 5.

① **Good**: a strengthener that strengthens the hypothesis or a weakener that weakened the hypothesis. For instance, it is rude to discuss people's weight problems, unless you are their nutritionist, then it is socially acceptable.

② **Neutral**: the update neither strengthened nor weakened the hypothesis. For example, the fact that the the debt is to the IRS doesn't change our perception about the extent that PersonX wants to become debt free.

③ **Weakener instead of strengthener**: the generated strengthener weakened the hypothesis.

④ **Strengthener instead of weakener**: the generated weakener strengthened the hypothesis.

⑤ **Restating the premise**: updates that roughly repeated the premise.

⑥ **Restating the hypothesis**: updates that roughly repeated the hypothesis.

⑦ **Contradicting the premise**: the generated update (implicitly or explicitly) contradicted the premise. For instance, when the premise mentions picking up someone at the airport, but the update talks about driving them there.

⑧ **Premise or hypothesis are nonsensical**: stemming from annotation errors in the original datasets.

⑨ **Update is nonsensical or other**: updates that are nonsensical.

Table 6 displays the percent of categories in each task and model. The results reconfirm the findings from the human evaluation in §6.2, with GPT2-XL leading with good generations with more than half of its generations for δ-SNLI and δ-SOCIAL judged as good. The Bart models suffer from various types of errors.

Dual-purpose updates. In addition, we looked into instances from the development where a single model generated an identical sentence as both a strengthener and weakener (for a given premise-hypothesis pair). Ideally, such instances should be rare, as a sentence may increase or decrease the likelihood of a hypothesis, but not both. In practice, we found such overlaps to be a very common failure mode. For a given premise-hypothesis input, we measure the frequency with which each model generates an identical sentence across the top five sampled strengtheners and top five sampled weakeners. The percentage of inputs resulting in such overlaps was extremely high for the Bart models: 96.53%, 97.53%, and 99.48% for δ-ATOMIC, δ-SOCIAL, and δ-SNLI, respectively (among 1900, 979, and 194 instances). The corresponding rates for the GPT2 models were much lower (although non-negligible): 48.42%, 33.91%, and 33.91%, respectively.

Is the Premise Necessary? In the classification task, we observe that models trained without access to the premise perform nearly as well as those trained with access to the full context (premise, hypothesis, update). This raises the interesting question of what role the premise plays in defeasible natural language inference. It is possible that in many cases, the premise is not as crucial as one might expect. Recall the classic example of defeasible reasoning: "Tweety is a bird" (premise), therefore "Tweety flies" (hypothesis), however "Tweety is a penguin" (update), and thus Tweety does not fly. In this case, it is evident that, while the premise was necessary to originally derive the hypothesis, the update alone is sufficient to conclude the hypothesis no longer holds.[5] In fact, the premise is entailed by the update, and perhaps even discernible from the hypothesis.

However, we should not conclude the premise is unnecessary in all cases. In the generative task, removing the premise makes only a slight difference in performance for δ-ATOMIC (Δ1.64%) but a substantial difference for δ-SNLI (Δ10.93%) (perhaps due to more specific contexts in SNLI). Because all generative models lag human performance, however, it may simply be a property of current models that they are unable to effectively leverage information from the premise; to match human performance, they may need to leverage this information.

For further analysis, we took outputs from the GPT2-XL $\mathcal{H}$-only model on SNLI and ask human evaluators to assess the outputs under two conditions: (1) annotator observing only the hypothesis, and (2) annotator observing both the premise and hypothesis. In 47.8% of cases, the output was labeled correct in both conditions; 34.1% of cases

[5]The question of the importance of the premise is perhaps relevant to another question that arose in earlier studies of defeasible inference, namely the role of human memory, and whether a belief could be defeated with new evidence if the holder of that belief did not recall the reason for it (Pollock, 1987).

were labeled incorrect in both conditions. Interestingly, in 12.4% of cases, the output was labeled correct in condition (1) and incorrect in condition (2). This finding points to a proportion of cases where the model would need to integrate information from the premise to generate valid strengtheners and weakeners.

8 Limitations of Elicitation

To collect the strengthener and weakener sentences in this work, we elicited sentences from crowdsource workers. Elicitation as a method of text data collection has a number of known flaws. In particular, (1) annotators may use label-dependent heuristics or strategies to produce sentences that introduce superficial correlations between text features and labels (Poliak et al., 2018; Gururangan et al., 2018; Tsuchiya, 2018); (2) elicitation may result in repeated responses of salient answers that are a small subset of all possible *valid* answers (McRae et al., 2005); and (3) elicited responses may contain implicit judgments or stereotypic associations about gender, race, and age, among others (Rudinger et al., 2017).

To avoid the first issue of annotation artifacts, we focus primarily on the *generative* task formulation, which is less susceptible to this problem. Regarding the second issue of coverage (or recall), we note that in this work we are concerned with whether it is possible for models to generate *any* correct weakeners or strengtheners in the first place; evaluating their ability to generate more exhaustively is a challenge we leave for future work. To address the third concern, we explicitly ask annotators to avoid such stereotyped associations in their responses. (See supplement for details.) This is an imperfect but expedient solution and for this reason we caution that the collected data is intended at this stage for scientific purposes only. Furthermore, we note that the elicited strengtheners and weakeners about social norms are subjective and, often, culturally dependent. This data should therefore be understood as *descriptive* of social norms (and their inherent subjectivity), rather than *prescriptive* of them.

9 Conclusion and Future Work

To the best of our knowledge, this is the first work to attempt merging long-standing ideas in AI about defeasible reasoning with contemporary formulations of natural language inference and commonsense reasoning tasks. We do this by crowdsourcing extensions to three existing inference datasets with enriched contexts that exemplify cases in which an inference is strengthened or weakened. From the collected data, we formulate a classification task and a generation task for defeasible inference in natural language. After demonstrating that the classification task is easily solved by state-of-the-art pretrained language models, we focus instead on the generative task of creating strengtheners or weakeners for a given premise-hypothesis pair, which we liken to "thinking like a skeptic." We demonstrate that fine-tuned language models successfully generate good-quality weakeners and strengtheners in 61-68% of cases.

Machine reasoning about the plausibility of inferences (Wang et al., 2018), let alone plausibility under different circumstances, is considered an unsolved problem and an obstacle to developing machine commonsense (Davis and Marcus, 2015). An inference engine with such capabilities may potentially be useful for various applications that require reassessing conclusions under changing conditions, such as processing legal texts (Hage, 2005) and mining arguments (Bilu and Slonim, 2016). In knowledge base completion, a "closed world" or default assumptions require the ability to defeat such assumptions given the appropriate counter evidence. Such ability was built into the Cyc inference engine (Lenat, 1995), but was largely absent from modern knowledge bases.

Yet, a number of challenges remain for future work. In our qualitative analysis of generated outputs (§7), we identify a number of systematic error types that future modeling efforts may seek to address. While this work addresses the quality and accuracy of generated outputs, we leave the more challenging task of evaluating the *coverage* (recall) of those outputs to future work. Finally, joint modeling between defeasible inference and related reasoning tasks such as abductive reasoning (Peirce, 1960; Bhagavatula et al., 2019) and counterfactual reasoning (Goodman, 1947; Qin et al., 2019; Tandon et al., 2019) is a potentially fruitful line of inquiry.

Acknowledgments

This work was supported by the Allen Institute for AI, the University of Washington, DARPA CwC through ARO (W911NF15-1-0543), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and DARPA KAIROS. The U.S. Gov-

ernment is authorized to reproduce and distribute reprints for governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsement.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530, Berlin, Germany. Association for Computational Linguistics.

Samuel 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.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of ACL*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin. Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nelson Goodman. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy*, 44(5):113–128.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

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.

Jaap Hage. 2005. Law and defeasibility. *Studies in legal logic*, pages 7–32.

Catherine Havasi, Robyn Speer, and Jason Alonso. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer.

Robert Koons. 2017. Defeasible reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, winter 2017 edition. Metaphysics Research Lab, Stanford University.

Sarit Kraus, Daniel Lehmann, and Menachem Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial intelligence*, 44(1-2):167–207.

Klaus Krippendorff. 1980. Content analysis: An introduction to its methodology.

Alex Lascarides and Nicholas Asher. 1991. Discourse relations and defeasible knowledge ‘. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 55–62.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and philosophy*, 16(5):437–493.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Hector J Levesque. 1990. All i know: a study in autoepistemic logic. *Artificial intelligence*, 42(2-3):263–309.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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*.

Bill MacCartney. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University.

John McCarthy. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial intelligence*, 13(1-2):27–39.

John McCarthy and Patrick J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press. Reprinted in McC90.

Drew McDermott and Jon Doyle. 1980. Non-monotonic logic i. *Artificial Intelligence*, 13(1):41 – 72. Special Issue on Non-Monotonic Logic.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. 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.

Charles Sanders Peirce. 1960. *Collected papers of charles sanders peirce*, volume 2. Harvard University Press.

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.

John L. Pollock. 1987. Defeasible reasoning. *Cognitive Science*, 11(4):481–518.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. -.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Raymond Reiter. 1980. A logic for default reasoning. *Artificial intelligence*, 13(1-2):81–132.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

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

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. 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 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. 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 303–308, New Orleans, Louisiana. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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

A Model Hyperparameters

Classification Task All models for the classification task were trained on a single NVIDIA Tesla P100 GPU on a Google Cloud instance. All models were fine-tuned with RoBERTa-base, which has 115M parameters. Best accuracy of five runs on the development set is reported in Table 7.

Task	Inputs	RoBERTa
δ-SNLI	$(\mathcal{P}, \mathcal{H}, \mathcal{U})$	83.3
	$(\varnothing, \mathcal{H}, \mathcal{U})$	81.1
	$(\varnothing, \varnothing, \mathcal{U})$	64.3
δ-ATOMIC	$(\mathcal{P}, \mathcal{H}, \mathcal{U})$	78.6
	$(\varnothing, \mathcal{H}, \mathcal{U})$	77.8
	$(\varnothing, \varnothing, \mathcal{U})$	65.7
δ-SOCIAL	$(\varnothing, \mathcal{H}, \mathcal{U})$	85.7
	$(\varnothing, \varnothing, \mathcal{U})$	72.0

Table 7: Accuracy (%) on the dev set of each classification baseline.

Generation Task All models were trained on a single NVIDIA Quadro RTX 8000 GPU. Runtime ranged between 5 minutes (GPT2-S on ATOMIC) to 3.5 hours (GPT2-XL on SNLI). The number of parameters ranges from 117M (GPT2-S) to 1.558B (GPT2-XL). Table 8 shows the generative models' performance on the dev set.

Task	LM	Macro Ppl.	Micro Ppl.	BLEU-4	ROUGE-L
δ-SNLI	GPT2-S	3.594	3.599	13.110	27.192
	GPT2-XL	2.838	2.842	17.963	**29.234**
	T5-large	8.617	8.632	12.849	25.962
	Bart-large	12.721	12.766	**18.289**	28.666
δ-ATOMIC	GPT2-S	3.178	3.178	3.739	18.232
	GPT2-XL	2.189	2.189	**4.595**	**21.345**
	T5-large	7.8	7.8	0.872	12.165
	Bart-large	10.174	10.174	4.083	20.335
δ-SOCIAL	GPT2-S	2.637	2.637	9.779	17.449
	GPT2-XL	2.176	2.176	**12.648**	19.616
	T5-large	5.754	5.755	4.486	15.123
	Bart-large	6.142	6.142	11.628	**19.350**

Table 8: Automatic evaluation results on the dev set, for the generative models.

B Crowdsourcing Task

Figures 3 and 4 display the full instructions shown to the crowdsourcing workers for the δ-SNLI and δ-ATOMIC update elicitation and for the social norms update elicitation, respectively.

SUMMARY
For this task, given a **PREMISE** and a **HYPOTHESIS** you will:
Write a **WEAKENER** in one sentence.
Write a **STRENGTHENER** in one sentence.

EXPLANATION
You will be presented with two sentences, called a **PREMISE** and a **HYPOTHESIS**, respectively.
The **PREMISE** sentence describes a **real-world situation** and is always assumed to be true.
The **HYPOTHESIS** sentence describes **an assumption or inference** that we might make about that situation having read the premise.
In most cases, the hypothesis statement is *very likely* to be true given the premise; however, it is not necessarily *guaranteed* to be true.

You will provide *additional information* about the situation that might **WEAKEN** or **STRENGTHEN** the hypothesis.
A **WEAKENER** is a statement that weakens the hypothesis;
it makes us much **less likely** to believe the hypothesis is true.
$\Longrightarrow$ **TIP**: Start by thinking of a situation where the **PREMISE** is true but the **HYPOTHESIS** is wrong.
A **STRENGTHENER** is a statement that strengthens the hypothesis;
it makes us much **more likely** to believe the hypothesis is true.
$\Longrightarrow$ **TIP**: start by thinking of a situation where both the **PREMISE** and the **HYPOTHESIS** are true.

EXAMPLES (omitted)

RULES
DO write in complete sentences.
DO use the "impossible" checkbox when appropriate! The use may be rare, but we do expect some.
For weakeners:
DO NOT contradict the premise
DO NOT simply negate the hypothesis
DO NOT directly contradict the hypothesis
For strengtheners:
DO NOT contradict the premise
DO NOT repeat or rephrase the contents of the hypothesis

IMPORTANT FINAL NOTE
The sentences you write here will be used in experiments to teach Artificial Intelligence (AI) systems how to make nuanced inferences about real-world situations usually involving people. *It is important that these AI systems treat all people fairly, regardless of their race, ethnicity, religion, gender, sexuality, ability, or other aspects of personal identity.* Therefore, when you perform this task, please bear this in mind.

Please AVOID writing sentences that involve making stereotyped or prejudicial assumptions about people based on their identities.
(Examples omitted)

If you absolutely cannot think of a strengthener or weakener that does not violate this guideline, then please select the "impossible" checkbox, followed by the appropriate selection. This option should be used very sparingly.

Figure 3: HIT Template for update elicitation for the δ-SNLI and δ-ATOMIC data.

SUMMARY
For this task, given a **GENERALIZATION** about a socially normative behavior or judgement:
Write an **UNDERMINING CONTEXT** in one sentence.
Write a **SUPPORTING CONTEXT** in one sentence.

EXPLANATION
You will be presented with one sentence, called a **GENERALIZATION**.
A **GENERALIZATION** is a statement that speaks of a socially normative behavior. In other words, it is a *generalizing* statement about how we expect people to behave in society.

You will provide *additional contexts* about the situation that might **UNDERMINE** or **SUPPORT** the generalization.
An **UNDERMINING** context provides a situation that weakens the generalization;
it makes the generalization **less** relevant or effective.
$\Longrightarrow$ **THINK**: "This generalization is makes sense **unless** _____"
A **SUPPORTING** provides a situation that strengthens the generalization;
it makes the generalization **more** relevant or effective.
$\Longrightarrow$ **THINK**: "This generalization is makes sense **especially if** _____"

EXAMPLES (omitted)

RULES
DO write in complete sentences.
DO use the "impossible" checkbox when appropriate! The use may be rare, but we do expect some.
For undermining context:
DO NOT contradict or negate the generalization
DO provide a real world situation that will weaken the generalization
For supporter:
DO NOT agree to or repeat the generalization
DO provide a real world situation that will strengthen the generalization

IMPORTANT FINAL NOTE
The sentences you write here will be used in experiments to teach Artificial Intelligence (AI) systems how to make nuanced inferences about real-world situations usually involving people. *It is important that these AI systems treat all people fairly, regardless of their race, ethnicity, religion, gender, sexuality, ability, or other aspects of personal identity.* Therefore, when you perform this task, please bear this in mind.

Please AVOID writing sentences that involve making stereotyped or prejudicial assumptions about people based on their identities.
(Examples omitted)

If you absolutely cannot think of a strengthener or weakener that does not violate this guideline, then please select the "impossible" checkbox, followed by the appropriate selection. This option should be used very sparingly.

Figure 4: HIT Template for update elicitation for the δ-SOCIAL data.