PInKS: Preconditioned Commonsense Inference with Weak Supervision

Anonymous ACL submission

Abstract

Reasoning with preconditions such as "glass can be used for drinking water unless the glass is shattered" remains an open-problem for language models. The main challenge lies in the scarcity of preconditions data and model's lack of support for such reasoning. We present *PInKS*, Preconditioned Commonsense Inference with WeaK Supervision, an improved model for reasoning with preconditions through minimum supervision. We show, both empirically and theoretically, that *PInKS* improves the results across the benchmarks on reasoning with the preconditions of commonsense knowledge (up to 0.4 Macro-F1 scores). We further investigate the robustness of our method through PAC-Bayesian informativeness analysis, recall measures and ablation study.

Figure 1: Examples on Preconditioned Inference and the NLI format they can be represented in.

1 Introduction

Inferring the effect of a situation or precondition on a subsequent action or state (illustrated in Fig. 1) is an open part of commonsense reasoning. It requires different dimensions of commonsense knowledge (Woodward, 2011), e.g. physical, causal, social, etc. This capability would improve many knowledge-driven tasks in question answering (Wang et al., 2019; Talmor et al., 2019), machine reading comprehension (Sakaguchi et al., 2020), and narrative prediction (Mostafazadeh et al., 2016). It will also benefit on a wide range of real-world intelligent applications such as legal document processing (Hage, 2005), claim verification (Nie et al., 2019) and debate processing (Widmoser et al., 2021).

Multiple recent studies have taken the effort on reasoning with preconditions of commonsense knowledge (Rudinger et al., 2020; Qasemi et al., 2021; Mostafazadeh et al., 2020; Hwang et al., 2020). These studies show that preconditioned reasoning represents an unresolved challenge to state-of-the-art (SOTA)language model (LM) based reasoners. Generally speaking, the problem of reasoning with preconditions has been formulated as variations of the natural language inference (NLI) task where, given a precondition/update, the model has to decide its effect on common sense statement or chain of statements. For example, CoreQuisite (Qasemi et al., 2021) approaches the task from causal (hard reasoning) perspective in term of *enabling* and *disabling* preconditions of commonsense knowledge, and evaluate reasoners with crowdsourced commonsense statements about the two polarities of preconditions of statements in ConceptNet (Speer et al., 2017). Similarly, $\delta-$NLI (Rudinger et al., 2020) formulates the problem in terms of soft assumptions, i.e., *weakeners* and *strengtheners*, and justify whether *update* sentences *weakens* or *strengthens* the textual entailment in sentence pairs from sources such as SNLI (Bowman et al., 2015). Obviously, both tasks capture the same phenomena of reasoning with preconditions and the slight difference in format does not hinder their usefulness (Gardner et al., 2019). As both works conclude, SOTA models fail to beat the task of reasoning with preconditions.

We identify two reasons for such shortcomings of LMs on reasoning with preconditions: 1) high cost to obtain sufficient training data, and 2) need of improved LMs to reason with such knowledge. First, current resources for preconditions of common sense are gathered through direct human supervision and crowdsourcing. First, current resources for preconditions of common sense are manually annotated. Although this yields highest quality of data, it is costly and not scalable. Second, off-the-shelf LMs are trained on unannotated corpora with no direct guidance on specific tasks. Although such models can be further fine-tuned to achieve impressive performance on a wide range of tasks, they are far from perfect in reasoning on preconditions due to their complexity of need for deep commonsense understanding and lack of large scale training data.

In this work, we present *PInKS* (see Fig. 2), a minimally supervised approach for reasoning with the precondition of commonsense knowledge in LMs. The main contributions are 3 points. **First**, to enhance training of the reasoning model (§3), we propose two strategies of retrieving rich amount of cheap weak supervision signals (Fig. 1). In the first strategy (§3.1), we use common linguistic patterns (e.g. "[action] unless [precondition]") to gather sentences describing preconditions and actions associated with them from massive free-text corpora (e.g. OMCS (Havasi et al., 2010)). The second strategy (§3.2) then uses generative data augmentation methods on top of the extracted sentences to induce even more training instances. As the **second** contribution (§3.3), we improve the LMs with more robust and generalized preconditioned commonsense inference. We modify the masked language model (MLM) learning objective to biased masking, which puts more emphasis on preconditions, hence improving the LMs capability to reason with preconditions. Finally, for **third** contribution, we go beyond empirical analysis of *PInKS* and investigate the performance and robustness through theoretical guarantees of PAC-Bayesian analysis (He et al., 2021).

Through extensive evaluation on five representative datasets (ATOMIC2020 (Hwang et al., 2020), WINOVENTI (Do and Pavlick, 2021), ANION (Jiang et al., 2021), *CoreQuisite* (Qasemi et al., 2021) and DNLI (Rudinger et al., 2020), we show that *PInKS* improves the performance of NLI models, up to 0.05 Macro-F1 without seeing any task-specific training data and up to 0.4 Macro-F1 in low-resource setup (§4.1). Here we use In addition to the empirical results, using theoretical guarantees of informativeness measure in *PABI* (He et al., 2021), we show (§4.2) that the minimally supervised data of *PInKS* is as informative as fully supervised datasets. Finally, to investigate the robustness of *PInKS*, we do ablation study of *PInKS* (§4.5)) and the effect of recall value of the noisy linguistic patterns used for *PInKS* (§4.4). Here, we study recall value's effect on the quality of the final model in terms of informativeness of the gathered data. The goal is study recall as proxy for precision to answer the question of "at what point does the noise in weakly supervised data become destructive?".

2 Problem Definition

Common sense statements describe well-known information about concepts, and, as such, they are acceptable by people without need for debate (Sap et al., 2019; Ilievski et al., 2020). The preconditions of common sense knowledge are eventualities that affect happening of a common sense statement (Hobbs, 2005). Here, we distinguish between possibility that a statement is true given preconditions and the general possibility that it is happening without extra information, however based on common sense they must both be agreed upon by humans. In this work

These preconditions can either *allow* or *prevent* the common sense statement (see Fig. 1) in different degrees (Rudinger et al., 2020; Qasemi et al., 2021). For example in some tasks the allowing or prevention conditions are modeled as strong constraints such as *enabling* and *disabling* (Qasemi et al., 2021), and others model soft constraints like *strengthening* and *weakening* (Rudinger et al., 2020). In addition, some tasks have strict constraint on the statement (Rudinger et al., 2020; Hwang et al., 2020) whereas others do not (Do and Pavlick, 2021; Qasemi et al., 2021). Using this definition of preconditions, then one way to formulate the problem of reasoning with them is as follows:

Definition 1 ***Preconditioned Inference:*** *given a common sense statement and an update sentence that serves as precondition, is the statement still allowed or prevented?*

This definition is consistent with definitions in previous works in the field, and serves as an unified definition to consolidate the literature. Here, similar to Rudinger et al. (2020), the update can have different levels of effect on the statement, from

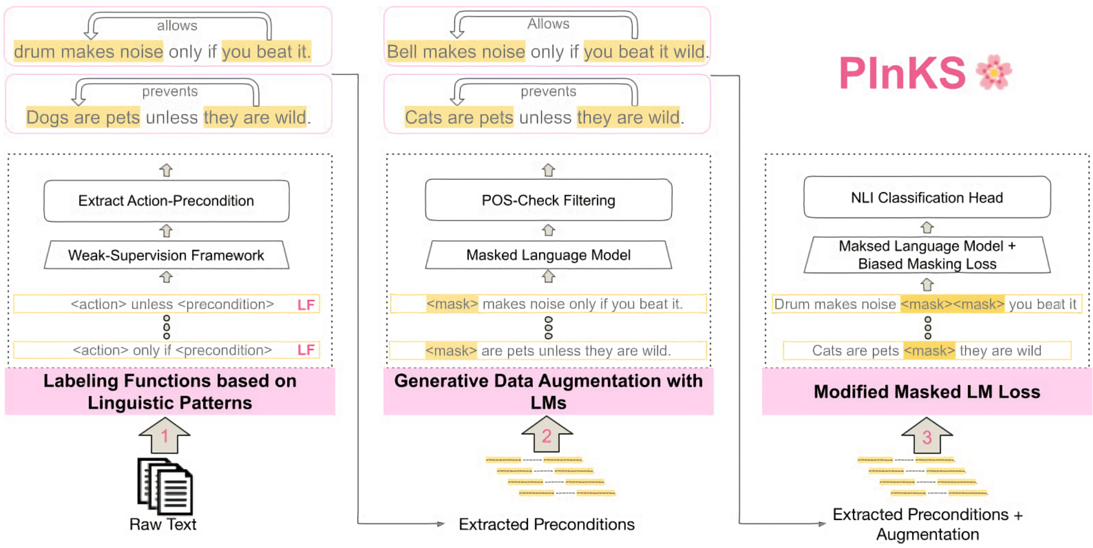

Figure 2: Overview of the three minimally supervised methods in *PInKS*.

causal connection (hard) to material implication (soft). In addition, similar to Qasemi et al. (2021), the statement can have any form and is not bound to the two-sentence structure in Rudinger et al. (2020).

3 Preconditioned Inference with Minimal Supervision

In *PInKS*, to overcome the challenges associated with inference with preconditions, we propose two sources of weak supervision to enhance the training of a reasoner: linguistic patterns to gather rich (but allowably noisy) preconditions (§3.1), and generative augmentation of the preconditions data (§3.2). The main hypothesis in using weak-supervision methods is that pre-training models on large amount of weakly labeled data could improve model's performance on similar downstream tasks (Ratner et al., 2017). In weak supervision terminology for heuristics, the experts design a set of heuristic labeling functions (LFs) that serves as the generators of the noisy label (Ratner et al., 2017). These labeling functions can produce overlapping or conflicting labels for a single instance of data that will need to be resolved either with simple methods such as ensemble inference or more sophisticated probabilistic methods such as data programming (Ratner et al., 2016), or generative (Bach et al., 2017). Here, the expert still needs to design the heuristics to query the knowledge and convert the results to appropriate labels for the task. In addition, we propose the modified language modeling objective that uses biased masking to improve the precondition-reasoning capabilities of LMs (§3.3).

3.1 Weak Supervision with Linguistic Patterns

We curate a large-scale automatically labeled dataset for, both type of, preconditions of commonsense statements by defining a set of linguistic patterns and searching through raw corpora. Finally, we have a post-processing filtering step to ensure the quality of the extracted preconditions.

Raw Text Corpora: In our experiments, we acquire weak supervision from two corpora: Open Mind Common Sense (OMCS) (Singh et al., 2002) and ASCENT (Nguyen et al., 2021a). OMCS is a large commonsense statement corpus that contains over 1M sentences from over 15,000 contributors. ASCENT has consolidated over 8.9M commonsense statements from the Web.

First, we use sentence tokenization in NLTK (Bird et al., 2009) to separate individual sentences in the raw text. Each sentence is then considered as a individual statement to be fed into the labeling functions. We further filter out the data instances based on the conjunctions used in the common sense statements after processing the labeling functions (discussed in Post-Processing paragraph).

Labeling Functions (LF): We design the LFs required for weak-supervision (discussed in §2), with a focus on the presence of a linguistic pattern in the sentences based on a conjunction (see Tab. 1 for examples). In this setup, each LF labels the training data as *Allowing*, *Preventing* or *Abstain*

Text	Label	Action	Precondition
A drum makes noise only if you beat it.	Allow	A drum makes noise	you beat it.
Your feet might come into contact with something if it is on the floor.	Allow	Your feet might come into contact with something	it is on the floor.
Pears will rot if not refrigerated	Prevent	Pears will rot	refrigerated
Swimming pools have cold water in the winter unless they are heated.	Prevent	Swimming pools have cold water in the winter	they are heated.

Table 1: Examples from the collected dataset through linguistic patterns in §3.1.

(no label assigned) depending on the linguistic pattern it is based on. For example, as shown in Tab. 1 the presence of conjunctions *only if* and *if*, with a specific pattern, suggests that the precondition *Allows* the action. Similarly, the presence of the conjunction *unless* indicates a *Preventing* precondition. We designed 20 such LFs based on individual conjunctions through manual inspection of the collected data in several iterations, for which details are described in appx. §A.1.

Extracting Action-Precondition Pairs Once the sentences have an assigned label, we extract the *action-precondition* pairs using the same linguistic patterns. This extraction can be achieved by leveraging the fact that a conjunction divides a sentence into *action* and *precondition* in the following pattern "*precondition conjunction action*", as shown in Tab. 1.

However, there are sentences in the dataset that contain multiple conjunctions. For instance, the sentence "Trees continue to grow for all their lives except in winter if they are not evergreen." includes two conjunctions "except" and "if". This occurrence of multiple conjunctions in a sentence leads to ambiguity in the extraction process. To overcome this challenge, we further make selection on the patterns by measuring their recalls. To do so, we sample 20 random sentences from each conjunction (400 total) and label them manually on whether they are relevant to our task or not by two expert annotators. If a sentence is relevant to the task, it is labeled as 1; otherwise, 0. We then average the score of two annotators for each pattern/conjunction to get its recall score. This recall score serves as an indicator of the quality of preconditions extracted by the pattern/conjunction in the context of our problem statement. Hence, priority is given to a conjunction with a higher recall in case of ambiguity. Further, we also set a minimum recall threshold (=0.7) to filter out the conjunctions having a low recall score (8 LFs), indicating low relevance to the task of reasoning with preconditions (see Appx. §A.1 for list of recall values).

Post-Processing On manual inspection of sentences matched by the patterns, we observed a few instances from random samples that were not relevant to the context of commonsense reasoning tasks, for example: *How do I know if he is sick?* or, *Pianos are large but entertaining*. We accordingly filter out sentences that are likely to be irrelevant instances. Specifically, those include 1) questions which are identified based on presence of question mark and interrogative words (List of interrogative words in Appx. §A.4), or 2) do not have a verb in their precondition. Through this process we end up with a total of 113,395 labeled action-precondition pairs with 102,474 *Allow* and 10,921 *Prevent* assertions.

3.2 Generative Data Augmentation

To further augment and diversify training data, we leverage another technique of retrieving weak supervision signals by probing LMs for generative data augmentation. To do so, we mask the nouns and adjectives from the text and let the generative language model fill in the masks with appropriate alternatives. The pivot-words here refers to the words in the text that are most responsible for giving meaning and context to the statement.

After masking the pivot-word and filling in the mask using LM, we filter out the augmentations that change the POS tag of the pivot-word and then keep the top 3 predictions for each mask. In addition, to keep the diversity of the augmented data, we do not use more than 20 augmented sentences for each original statement (picked randomly). For example, in the statement "Dogs are pets unless they are wild", the pivot-words are "dogs", "pets" and "wild". Upon masking "dogs", using RoBERTa (large) language model, we get valid augmentations such as "Cats are pets unless they are wild". Using this generative data augmentation, we end up with $7M$ labeled action-precondition pair with 11% *prevent* preconditions.

3.3 Precondition-Aware Biased Masking

To increase LMs' emphasis on preconditions, we used biased masking on conjunctions as the closest proxies to preconditions' reasoning. Based on this observation, we devised a biased masked language modeling loss that solely focuses on masking conjunctions in the sentences instead of random tokens. Similar to Dai et al. (2019), we mask the whole conjunction word in the sentence and ask the LM to fulfill the mask. The goal here is to start from a pretrained language model and, through this additional fine-tuning step, improve its ability to reason with preconditions. To use such fine-tuned LM in a NLI module, we further fine-tune the "LM+classification head" on subset of MNLI (Williams et al., 2018) dataset. Later in §4.5 we provide ablation study to showcase effectiveness of this additional fine-tuning step. For full list of conjunctions and implementation details check Appx. §A.3.

4 Experiments

This section, first showcases improvements of *PInKS* on five representative tasks for preconditioned inference (§4.1), theoretically backs the improvements using *PABI* (He et al., 2021) score (§4.2), and investigate a different fine-tuning strategy (§4.3). We then experiment on the effect of recall (discussed in §3.1) on *PInKS* using *PABI* score (§4.4). Finally, we do ablation study to evaluate effect of each step in *PInKS* (§4.5).

4.1 Evaluation on Target Tasks

Comparing the capability for models to reason with preconditions across different tasks (datasets) requires that inputs and outputs in such tasks be in the same canonical format. We used natural language inference (NLI) as such a canonical format. *CoreQuisite* (Qasemi et al., 2021) and δ-NLI (Rudinger et al., 2020) are already in NLI format and others can be converted easily using the groundwork laid in Qasemi et al. (2021). In NLI, given a sentence pair with a *hypothesis* and a *premise*, one predicts whether the hypothesis is true (entailment), false (contradiction), or undetermined (neutral) given the premise (Williams et al., 2018). Each task is preserved with equivalence before and after any format conversion at here, hence conversion does not seek to affect the task performance, inasmuch as it is discussed by Gardner et al. (2019). More details on this conversion process are in Appx. §B, and examples from the original target datasets are given in Tab. 10.

Setup To implement and execute labeling functions and resolve labeling conflict, we use Snorkel (Ratner et al., 2017), one of the SOTA frameworks for algorithmic labeling on raw data that provides ease-of-use APIs.[1] For more details on Snorkel and its setup details, please see Appendix A.2.

For each target task, we start from a pretrained model (RoBERTa-Large-MNLI (Liu et al., 2019)), fine-tune it on *PInKS* and evaluate its performance on the test portion of the target dataset in two setups: zero-shot transfer learning(w.r.t. target dataset; labeled as *PInKS* column) and fine-tuned on the training portion of the target task (labeled as *Orig.+PInKS*). To facilitate comparison, we also provide the results for fully fine-tuning on the training portion of the target task and evaluating on its testing portion (labeled as *Orig.* column; no *PInKS* is used here). To create the test set, if the original data does not provide a split (e.g. ATOMIC and Winoventi), we use unified random sampling with the $[0.45, 0.15, 0.40]$ ratio for train/dev/test. The experiments are conducted on a commodity workstation with an Intel Xeon Gold 5217 CPU and an NVIDIA RTX 8000 GPU. For all the tasks, we used the implementation, and pretrained weights from *huggingface* (Wolf et al., 2020) and utilized PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) library to manage the fine-tuning process. We evaluate each performance by aggregating the *Macro-F1* score (implemented in Pedregosa et al. (2011)) on the ground-truth labels and report the results on the unseen test split of the data.

Target Data	Orig.	*PInKS*	Orig+*PInKS*
δ-NLI	83.4	60.3	84.1
CoreQuisite	77.1	69.5	68.0
ANION	81.1	52.9	81.2
ATOMIC	43.2	48.0	88.6
Winoventi	51.1	52.4	51.0

Table 2: Macro-F1 (%) results of *PInKS* on the target datasets: no *PInKS* (*Orig.*), with *PInKS* in zero-shot transfer learning setup (*PInKS*) and *PInKS* in addition to original task's data (*Orig.+PInKS*)

Discussion Table 2 summarizes the evaluation results of this section. As illustrated, *PInKS*

[1]Other alternatives such as skweak (Lison et al., 2021) can also be used for this process.

can achieve on-par results with the direct supervision from the task-specific training data. On ATOMIC (Hwang et al., 2020) and Winoventi (Do and Pavlick, 2021), *PInKS* exceeds the supervised results even without seeing any examples from the target data (zero-shot transfer learning setup). On δ-NLI (Rudinger et al., 2020), ANION (Jiang et al., 2021) and ATOMIC (Hwang et al., 2020), combination of *PInKS* and train subset of target task (*PInKS* in low-resource setup) outperforms the target task results. This shows *PInKS* can also utilize additional data from target task to achieve better performance consistently across different aspects of preconditioned inference. However, on *CoreQuisite* (Qasemi et al., 2021), *PInKS* is not able to outperform original target task results in none of the setups. This can be attributed to nature of data in *CoreQuisite* in which contrary to other tasks focuses on hard preconditions instead of soft ones. This result is also consistent with their results on transfer learning from soft to hard preconditioned reasoning.

4.2 Informativeness Measure

PABI (He et al., 2021) proposes a unified PAC-Bayesian motivated informativeness measure that correlates with the improvements provided by the incidental signals to showcase its effectiveness on target task. The incidental signal can include an inductive signal, e.g. partial/noisy labeled data, or a transductive signal, e.g. cross-domain signal in transfer learning. In this experiment, we go beyond the empirical results and use the *PABI* measure to showcase how improvements from *PInKS* are theoretically backed.

Setup We carry over the setup on models and tasks from §4.1. For details on the *PABI* itself and the measurement details associated with it, please see Appx. §D.

Discussion Tab. 3 summarizes the *PABI* informativeness measure. Here the *PInKS* is compared with the rest of the dataset when considered as incidental signal, while considering δ-NLI (Rudinger et al., 2020) and CoreQuisit (Qasemi et al., 2021) as target tasks. Here although, *PInKS* is not the top informative incidental signal on the target dataset, its *PABI* numbers are still significant considering that its weak-supervision data are automatically acquired, while others are acquired based on human effort.

Indirect Data	*PABI* on *CoreQuisite*	*PABI* on δ-NLI
PInKS	36.6	19.1
δ-NLI	52.2	**85.5**
CoreQuisite	**52.3**	31.3
ANION	34.1	13.9
ATOMIC	20.9	17.4
Winoventi	36.4	53.4

Table 3: *PABI* informativeness measures (x100) of *PInKS* and other target tasks w.r.t *CoreQuisite* and δ-NLI. Bold values represent the maximum achievable *PABI* Score by considering train subset as *indirect* signal for test subset of respective data.

4.3 Curriculum vs. Multitask Learning

For results of §4.1, we considered the target task and *PInKS* as separate datasets, and fine-tuned model sequentially on them (curriculum learning;Pentina et al., 2015). We chose *curriculum* learning setup due to its simplicity in implementation, ease of fine-tuning process monitoring and hyperparameter setup. It would also allow us to look and the each task separately that increases interpretability of results.

However, in an alternative fine-tuning setup, one can merge the two datasets into one and fine-tune the model on the aggregate dataset (multi-task learning;Caruana, 1997). Here, we investigate such alternative and its effect on the results of §4.1.

Setup We use the same setup as §4.1 for fine-tuning the model on *Orig.+PInKS*. Here instead of first creating *PInKS* and then fine-tuning it on the target task, we merge the weakly supervised data of *PInKS* with the training subset of the target task and then do fine-tuning on the aggregate dataset. To manage length of this section, we only consider *CoreQuisite*, δ-NLI and Winoventi as the target dataset.

Target Data	Orig+*PInKS* (Multi-Task)	Diff.
δ-NLI	72.1	-11.00
CoreQuisite	77.3	+9.3
Winoventi	51.7	+0.7

Table 4: Macro-F1 (x100) results of *PInKS* on the target datasets using *multi-task* fine-tuning strategy and its difference with *curriculum* strategy.

Discussion Tab. 4 summarizes the results for *multi-task* learning setup and its difference w.r.t to the results of the *curriculum* learning setup in Tab. 2. Using *multi-task* learning does not show the consistent result across tasks. We see significant performance loss on δ-NLI on one hand and major performance improvements on *CoreQuisite* on the other. The Winoventi, however appears to not change as much in the new setup. We leave further analysis of *curriculum learning* to future work.

4.4 Informativeness vs. Recall

As mentioned in §3.1, each linguistic pattern is assigned a recall value calculated from expert annotations on its matches. Using this recall value coupled with the *PABI* informativeness measure, we can investigate the effect of the linguistic pattern's recal on quality of the extracted data.

Setup The model setup in this section is the same as the §4.1 and §4.4. Here, create different versions of *PInKS* with different levels of recall threshold (0.0, 0.5) and compare their informativeness on *CoreQuisite* (Qasemi et al., 2021) with *PInKS*'s (recall 0.75) informativeness. Here, to limit the computation time, we only use $100K$ samples from *PInKS* in each threshold value, which is especially important in lower thresholds due to huge size of extracted patterns with low recall threshold.

Discussion Tab. 5 summarizes the *PABI* informativeness estimation on weak supervision data under three threshold levels of recall on, and compare them with "zero rate" classifier (always predicting major class). As illustrated, the informativeness show a significant drop in lower recall showcasing the importance of using high recall templates in our weak-supervision task. For higher thresholds (0.95) the data will mostly consist of *allow* patterns, the model drops to near zero rate informativeness baseline. This susceptibility on pattern recall can be mitigated with having more fine-grained patterns on larger corpora. We leave further analysis on recall of patterns to future work.

Source Data	*PABI* on *CoreQuisite*
Zero Rate	25.5
PInKS-recall-0.00	23.8
PInKS-recall-0.50	25.6
PInKS-recall-0.70	36.6
PInKS-recall-0.95	26.2

Table 5: *PABI* informativeness measures of *PInKS* with different recall thresholds on *CoreQuisite*.

4.5 Ablation Study

As a final study, we focus on different aspects of *PInKS* and evaluate how each step is contributing to the results. There are three questions that needs to be addressed. First, how each labeling function (LF) is contributing to the extracted preconditions? Second, to what extend the weak supervision data contribute? (addressed in §4.1) And third, how much does the precondition-aware masking (§3.3) effect the overall performance of *PInKS*. Here, we try to address these question.

LF Analysis To address first question, we use statistics generated by Snorkel on top performing LFs (see Tab. 6). We study Coverage (fraction of raw corpus instances covered by the labeling function), Overlaps (fraction of raw corpus instances with at least two non-abstain labels), and Conflicts (fraction of the raw corpus instances with conflicting (non-abstain) labels) on top performing LFs. Here the polarity column refers to the non-abstain label that each LF can generate (all can output abstain as label).

Conjunction	Pol.	Coverage	Overlaps	Conflicts
in case	[1]	0.000227	0.000001	$6.19\text{x}10^{-7}$
to understand event	[1]	0.009189	0.000005	$4.64\text{x}10^{-6}$
statement is true	[1]	0.001647	0.000004	$4.12\text{x}10^{-6}$
except	[0]	0.000753	0.000003	$2.06\text{x}10^{-6}$
unless	[0]	0.000745	0.000007	$5.88\text{x}10^{-6}$
if not	[0]	0.000156	0.000002	$1.44\text{x}10^{-6}$

Table 6: Statistical analysis of labeling functions on raw data instances.

Effectiveness of Biased Masking For the third question, we focus on *CoreQuisite* as the target task and compare the results of *PInKS* with an alternative setup with no biased masking. In the alternative setup, we only use the weakly-supervised data that we extract to fine-tune RoBERTa-Large-MNLI model and compare the results. Our results show that the Macro-F1 score for zero-shot transfer learning setup drops to 68.4% from 69.5% without biased masking process.

5 Related Work

Reasoning with Preconditions The problem of collecting preconditions of common sense and reasoning with them has been studied in multiple works. Rudinger et al. (2020) uses the notion of "defeasible inference" (Pollock, 1987; Levesque, 1990) in term of how a new piece of information (*update*) *weakens* or *strengthens* a common sense hypothesis statement in relation to a premise sentence. For example, given the premise "Two men and a dog are standing among rolling green hills.", the knowledge that "The men are studying a tour map" weakens the hypothesis that "they are farmers", whereas "The dog is a sheep dog" strengthens it. Similarly, CoreQuisite (Qasemi et al., 2021) uses the notion of "causal complex" from Hobbs (2005), and defines preconditions as eventualities that either *allow* or *prevent* (allow negation (Fikes and Nilsson, 1971) of) a common sense statement to happen. For example, for the knowledge "the glass is shattered" prevents the statement "A glass is used for drinking water", whereas "there is gravity" allows it. In CoreQuisite, based on Shoham (1990) and Hobbs (2005), authors distinguish between two type of preconditions, causal connections (*hard*), and material implication (tends to cause; *soft*). As mentioned in §2, our definition covers these definitions and is consistent with both.

Hwang et al. (2020), Sap et al. (2019), Heindorf et al. (2020), and Speer et al. (2017), provided representations for preconditions of statements in term of relation types, e.g. *xNeed* in ATOMIC2020 (Hwang et al., 2020). However, the focus in none of these works is on evaluating SOTA models on such data. The closest study of preconditions to our work are Rudinger et al. (2020), Qasemi et al. (2021), Do and Pavlick (2021) and Jiang et al. (2021). In these works, direct human supervision (crowdsourcing) is used to gather preconditions of commonsense knowledge and they all show the shortcomings of SOTA models on comprehending with such knowledge. Our work differs as we rely on combination of distant-supervision and targeted fine-tuning instead of direct supervision to achieve on-par performance. Similarly, Mostafazadeh et al. (2020), and Kwon et al. (2020) also study the problem of reasoning with preconditions. However they do not explore *preventing* preconditions.

Weak Supervision In weak-supervision, the objective is similar to supervised learning. However instead of using human/expert resource to directly annotate unlabeled data, one can use the experts to design user-defined patterns to infer "noisy" or "imperfect" labels (Rekatsinas et al., 2017; Zhang et al., 2017; Dehghani et al., 2017), e.g. using heuristic rules. In addition, other methods such as repurposing of external knowledge (Alfonseca et al., 2012; Bunescu and Mooney, 2007; Mintz et al., 2009) or other types of domain knowledge (Stewart and Ermon, 2017) also lie in the same category. Weak supervision has been used extensively in NLU. For instance, Zhou et al. (2020) utilize weak-supervision to extract temporal commonsense data from raw text, Brahman et al. (2020) use it to generate reasoning rationale, Dehghani et al. (2017) use it for improved neural ranking models, and Hedderich et al. (2020) use it to improve translation in African languages. Similar to our work, ASER (Zhang et al., 2020) and ASCENT (Nguyen et al., 2021b) use weak supervision to extract relations from unstructured text. However, do not explore preconditions and cannot express *preventing* preconditions. As they do focus on reasoning evaluation, the extent in which their contextual edges express *allowing* preconditions is unclear.

Generative Data Augmentation Language models can be viewed as knowledge bases that implicitly store vast knowledge on the world. Hence querying them as a source of weak-supervision is a viable approach. Similar to our work, Wang et al. (2021) use LM-based augmentation for saliency of data in tables, Meng et al. (2021) use it as a source of weak-supervision in named entity recognition, and Dai et al. (2021) use masked LMs for weak supervision in entity typing.

6 Conclusion

In this work we presented *PInKS*, as an improved method for preconditioned commonsense reasoning which involves two techniques of weak supervision. To maximize the effect of the weak supervision data, we modified the masked language modeling loss function using biased masking method to put more emphasis on conjunctions as closest proxy to preconditions. Through empirical and theoretical analysis of *PInKS*, we show it significantly improves the results across the benchmarks on reasoning with the preconditions of commonsense knowledge. In addition, we show the results are robust in different recall values using the *PABI* informativeness measure and extensive ablation study.

Future work can consider improving the robustness of preconditioned inference models using methods such as virtual adversarial training (Miyato et al., 2018; Li and Qiu, 2020). With advent of visual-language models such as Li et al. (2019), preconditioned inference should also expand beyond language and include different modalities (such as image or audio). To integrate in down-steam tasks, one direction is to include such models in aiding inference in the neuro-symbolic reasoners, e.g. Lin et al. (2019); Verga et al. (2020).

Ethical Consideration

We started from openly available data that is both crowdsource-contributed and neutralized, however they still may reflect human biases. For example in case of *CoreQuisite* (Qasemi et al., 2021) they use ConceptNet as source of commonsense statements which multiple studies have shown in bias and ethical issues, e.g. (Mehrabi et al., 2021).

During design of labeling functions we did not collect any sensitive information and the corpora we used were both publicly available however they can also contain various types of bias. The labeling functions in *PInKS* are only limited to English language patterns, which may additional cultural bias to the data. However, our expert annotators did not notice any offensive language in data or the extracted preconditions.

Given the urgency of addressing climate change we have reported the detailed model sizes and runtime associated with all the experiments in Appendix C.

References

Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido. 2012. Pattern learning for relation extraction with a hierarchical topic model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, page 54–59, USA. Association for Computational Linguistics.

Stephen H Bach, Bryan He, Alexander Ratner, and Christopher Ré. 2017. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, pages 273–282. PMLR.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642, Lisbon, Portugal.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2020. Learning to rationalize for nonmonotonic reasoning with distant supervision. *arXiv preprint arXiv:2012.08012*.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 576–583.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-fine entity typing with weak supervision from a masked language model. *arXiv preprint arXiv:2106.04098*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Nam Do and Ellie Pavlick. 2021. Are rotten apples edible? challenging commonsense inference ability with exceptions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2061–2073.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Richard E Fikes and Nils J Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. *AIJ*, 2(3-4):189–208.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*.

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

Catherine Havasi, Robert Speer, Kenneth Arnold, Henry Lieberman, Jason Alonso, and Jesse Moeller. 2010. Open mind common sense: Crowd-sourcing for common sense. In *Workshops at the Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Hangfeng He, Mingyuan Zhang, Qiang Ning, and Dan Roth. 2021. Foreseeing the Benefits of Incidental Supervision. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Michael A Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on african languages. *arXiv preprint arXiv:2010.03179*.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM*, pages 3023–3030.

Jerry R Hobbs. 2005. Toward a useful concept of causality for lexical semantics. *Journal of Semantics*, 22(2):181–209.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Filip Ilievski, Pedro Szekely, and Daniel Schwabe. 2020. Commonsense knowledge in wikidata. In *ISWC Wikidata workshop*.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. " i'm not mad": Commonsense implications of negation and contradiction. *arXiv preprint arXiv:2104.06511*.

Heeyoung Kwon, Mahnaz Koupaee, Pratyush Singh, Gargi Sawhney, Anmol Shukla, Keerthi Kumar Kallur, Nathanael Chambers, and Niranjan Balasubramanian. 2020. Modeling preconditions in text with a crowd-sourced dataset. In *EMNLP-Findings*, pages 3818–3828, Online.

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

Linyang Li and Xipeng Qiu. 2020. Tavat: Token-aware virtual adversarial training for language understanding. *arXiv preprint arXiv:2004.14543*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP*, pages 2829–2839, Hong Kong, China.

Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. 2021. skweak: Weak supervision made easy for nlp. *arXiv preprint arXiv:2104.09683*.

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*.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. *arXiv preprint arXiv:2103.11320*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10367–10378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL-HLT*, pages 839–849, San Diego, California.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *EMNLP*, pages 4569–4586, Online.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021a. Advanced semantics for commonsense knowledge extraction. In *Proceedings of the Web Conference 2021*, pages 2636–2647.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021b. Advanced semantics for commonsense knowledge extraction. In *WWW*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830.

Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. 2015. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5492–5500.

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

Ehsan Qasemi, Filip Ilievski, Muhao Chen, and Pedro Szekely. 2021. Corequisite: Circumstantial preconditions of common sense knowledge. *arXiv preprint arXiv:2104.08712*.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29:3567–3575.

Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. 2017. Holoclean: Holistic data repairs with probabilistic inference. *arXiv preprint arXiv:1702.00820*.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *EMNLP-Findings*, pages 4661–4675, Online.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, volume 34, pages 8732–8740.

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 *AAAI*, pages 3027–3035.

Yoav Shoham. 1990. Nonmonotonic reasoning and causation. *Cognitive Science*, 14(2):213–252.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Russell Stewart and Stefano Ermon. 2017. Label-free supervision of neural networks with physics and domain knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*, pages 4149–4158, Minneapolis, Minnesota.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William W Cohen. 2020. Facts as experts: Adaptable and interpretable neural memory over symbolic knowledge. *arXiv preprint arXiv:2007.00849*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, pages 3261–3275.

Fei Wang, Kexuan Sun, Jay Pujara, Pedro Szekely, and Muhao Chen. 2021. Table-based fact verification with salience-aware learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4025–4036, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Manuel Widmoser, Maria Leonor Pacheco, Jean Honorio, and Dan Goldwasser. 2021. Randomized deep structured prediction for discourse-level processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1174–1184, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *ACL*, pages 1112–1122, New Orleans, Louisiana.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, pages 38–45, Online.

James Woodward. 2011. Psychological studies of causal and counterfactual reasoning. *Understanding counterfactuals, understanding causation. Issues in philosophy and psychology*, pages 16–53.

Ce Zhang, Christopher Ré, Michael Cafarella, Christopher De Sa, Alex Ratner, Jaeho Shin, Feiran Wang, and Sen Wu. 2017. Deepdive: Declarative knowledge base construction. *Communications of the ACM*, 60(5):93–102.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW*, pages 201–211. ACM / IW3C2.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. *arXiv preprint arXiv:2005.04304*.

A Details on *PInKS* Method

In this section, we discuss some of the extra details related to *PInKS* and its implementation.

A.1 Linguistic Patterns for *PInKS*

We use a set of conjunctions to extract sentences that follow the action-precondition sentence structure. Initially, we started with two simple conjunctions-*if* and *unless*, for extracting assertions containing *Allowing* and *Preventing* preconditions, respectively. To further include similar sentences, we expanded our vocabulary by considering the synonyms of our initial conjunctions. Adding the synonyms of *unless* we got the following set of new conjunctions for *Preventing* preconditions-{*but, except, except for, if not, lest, unless*}, similarly we expanded the conjunctions for Enabling preconditions using the synonyms of *if*-{*contingent upon, in case, in the case that, in the event, on condition, on the assumption, supposing*}. Moreover, on manual inspection of the OMCS and ASCENT datasets, we found the following conjunctions that follow the Enabling precondition sentence pattern-{*makes possible, statement is true, to understand event*}. Tab. 7, summarizes the final patterns used in *PInKS*, coupled with their recall value and their associated conjunction.

A.2 Details of Snorkel Setup

Beyond a simple API to handle implementing patterns and applying them to the data, Snorkel's main purpose is to model and integrate noisy signals contributed by the labeling functions modeled as noisy, independent voters, which commit mistakes uncorrelated with other LFs.

To improve the predictive performance of the model, Snorkel additionally models statistical relationships between LFs. For instance, the model takes into account similar heuristics expressed by two LFs to avoid "double counting" of voters. Snorkel, further, models the generative learner as a factor graph. A labeling matrix Λ is constructed by applying the LFs to unlabeled data points. Here, $\Lambda_{i,j}$ indicates the label assigned by the j^{th} LF for the i^{th} data point. Using this information, the generative model is fed signals via three factor types, representing the labeling propensity, accuracy, and pairwise correlations of LFs.

$$\phi_{i,j}^{Lab}(\Lambda) = \mathbb{1}\{\Lambda_{i,j} \neq \emptyset\}$$
$$\phi_{i,j}^{Acc}(\Lambda) = \mathbb{1}\{\Lambda_{i,j} = y_i\}$$
$$\phi_{i,j,k}^{Corr}(\Lambda) = \mathbb{1}\{\Lambda_{i,j} = \Lambda_{i,k}\}$$

The above three factors are concatenated along with the potential correlations existing between the LFs and are further fed to a generative model which minimizes the negative log marginal likelihood given the observed label matrix Λ.

A.3 Modified Masked Language Modeling

Tab. 8 summarizes the list of *Allowing* and *Preventing* conjunctions which the modified language modeling loss function is acting upon.

A.4 Interrogative Words

On manual inspection of the dataset, we observed some sentences that were not relevant to the common sense reasoning task. Many of such instances were interrogative statements. We filter out such cases based on the presence of interrogative words in the beginning of a sentence. These interrogative words are listed below.

Interrogative words: ["Who", "What", "When", "Where", "Why", "How", "Is", "Can", "Does", "Do"]

B Details on Target Data Experiments

For converting Rudinger et al. (2020), similar to Qasemi et al. (2021), we concatenate the "Hypothesis" and "Premise" and consider then as NLI's hypothesis. We then use the "Update" sentence as NLI's premise. The labels are directly traslated based on *Update* sentences's label, *weakener* to *prevent* and the *strengthener* to *allow*.

To convert the ATOMIC2020 (Hwang et al., 2020), similar to Qasemi et al. (2021), we focused on three relations *HinderedBy*, *Causes*, and *xNeed*. From these relations, edges with *HinderedBy* are converted as *prevent* and the rest are converted as *allow*.

Winoventi (Do and Pavlick, 2021), proposes Winograd-style entailment schemas focusing on negation in common sense. To convert it to NLI style, we first separate the two sentences in the *masked_prompt* of each instance to form *hypothesis* and *premise*. We get two versions of *premise* by replacing the MASK token in *premise* with their *target* or *incorrect* tokens. For the labels the version with *target* token is considered as *allow* and the version with *incorrect* token as *prevent*.

ANION (Jiang et al., 2021), focuses on contradiction in general. We focus on their commonsense contradiction subset as it is clean of lexical hints. Then we convert their crowdsourced *original head*

Conjunctions	Recall	Pattern
but	0.17	{action} but {negative_precondition}
contingent upon	0.6	{action} contingent upon {precondition}
except	0.7	{action} except {precondition}
except for	0.57	{action} except for {precondition}
if	0.52	{action} if {precondition}
if not	0.97	{action} if not {precondition}
in case	0.75	{action} in case {precondition}
in the case that	0.30	{action} in the case that {precondition}
in the event	0.3	{action} in the event {precondition}
lest	0.06	{action} lest {precondition}
makes possible	0.81	{precondition} makes {action} possible.
on condition	0.6	{action} on condition {precondition}
on the assumption	0.44	{action} on the assumption {precondition}
statement is true	1.0	The statement "{event}" is true because {precondition}.
supposing	0.07	{action} supposing {precondition}
to understand event	0.87	To understand the event "{event}", it is important to know that {precondition}.
unless	1.0	{action} unless {precondition}
with the proviso	-	{action} with the proviso {precondition}
on these terms	-	{action} on these terms {precondition}
only if	-	{action} only if {precondition}
make possible	-	{precondition} makes {action} possible.
without	-	{action} without {precondition}
excepting that	-	{action} excepting that {precondition}

Table 7: Linguistic patterns in *PInKS* and their recall value. For patterns with not enough match in the corpora have empty recall values.

Type	Conjunctions
Allowing	only if, subject to, in case, contingent upon, given, if, in the case that, in case, in the case that, in the event, on condition, on the assumption, only if, so, hence, consequently, on these terms, subject to, supposing, with the proviso, so, thus, accordingly, therefore, as a result, because of that, as a consequence, as a result
Preventing	but, except, except for, excepting that, if not, lest, saving, without, unless

Table 8: List of conjunctions used in modified masked loss function in section 3.3

or *contradiction head* as hypothesis, and the lexicalized predicate and tail as the premise (e.g. *xIntent* to *PersonX intends to*). Finally the label depends on head is *allow* for *original head* and *prevent* for *contradiction head*. We also replace "PersonX" and "PersonY" with random human names (e.g. "ALice", "Bob").

Finally, for the CoreQuisite (Qasemi et al., 2021), we used their proposed P-NLI task as a NLI-style task derived from their preconditions dataset. We converted their *Disabling* and *Enabling* labels to *prevent* and *allow* respectively.

Tab. 10 summarizes the conversion process through examples from the original data and the NLI task derived from each.

C Model Sizes and Run-times

For all the fine-tuning results in Tab. 2, Tab. 3 we used "RoBERTa-Large-MNLI" with 356M tuneable parameters. The mean run-time on target datasets is 1hr 55mins.

For the augmentation in *PInKS* dataset, we used "BERT" language model with $234M$ tuneable parameters. The mean run-time on the extracted sentences is 49hr.

D Details on *PABI* Measurement

PABI provides an Informativeness measure that quantifies the reduction in uncertainty provided by incidental supervision signals. We use the *PABI* measure to study the impact of transductive cross-domain signals obtained from our weakly-

Conjunction	Pol.	Pattern
to understand event	[1]	To understand the event "{event}", it is important to know that {precondition}.
in case	[1]	{action} in case {precondition}
statement is true	[1]	The statement "{event}" is true because {precondition}.
except	[0]	{action} except {precondition}
unless	[0]	{action} unless {precondition}
if not	[0]	{action} if not {precondition}

Table 9: Filtered Labeling Functions Patterns and their associated polarity.

supervised approach.

Following (He et al., 2021), in order to calculate *PABI* $\hat{S}(\pi_0, \tilde{\pi}_0)$, we first find out η, the difference between a perfect system and a gold system in the target domain $\mathcal{D}$ that uses a label set $\mathcal{L}$ for a task, using Eq.1.

$$\begin{aligned}\eta &= \mathbb{E}_{x \sim P_{\mathcal{D}(x)}} 1(c(x) \neq \tilde{c}(x)) \\ &= \frac{(|\mathcal{L}| - 1)(\acute{\eta_1} - \eta_2)}{1 - |\mathcal{L}|(1 - \acute{\eta_1})} \\ &= \frac{(|\mathcal{L}| - 1)(\eta_1 - \eta_2)}{1 - |\mathcal{L}|(1 - \eta_1)}\end{aligned} \tag{1}$$

Here, $P_{\mathcal{D}(x)}$ indicates the marginal distribution of x under $\mathcal{D}$, $c(x)$ refers to gold system on gold signals, $\tilde{c}(x)$ is a perfect system on incidental signals, η_1 refers to the difference between the silver system and the perfect system in the source domain, $\acute{\eta_1}$ indicates difference between the silver system and the perfect system in the target domain, and η_2 is the difference between the silver system and the gold system in the target domain.

Using Eq.1, the informative measure supplied by the transductive signals can be calculated as $\hat{S}(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{\eta \ln(|\mathcal{L}|-1) - \eta \ln \eta - (1-\eta)\ln(1-\eta))}{\ln|\mathcal{L}|}}$.

Name	Original Data		Derived NLI	
Winoventi (Do and Pavlick, 2021)	**masked_prompt**:	Margaret smelled her bottle of maple syrup and it was sweet. The syrup is {MASK}.	**Hypothesis**:	Margaret smelled her bottle of maple syrup and it was sweet.
	target:	edible	**Premise**:	The syrup is edible/malodorous
	incorrect:	malodorous	**Label**:	ENTAILMENT/CONTRADICTION
ANION (Jiang et al., 2021)	**Orig_Head**:	PersonX expresses PersonX's delight.	**Hypothesis**:	Alice expresses Alice's delight/anger.
	Relation:	xEffect	**Premise**:	feel happy.
	Tail:	Alice feel happy	**Label**:	ENTAILMENT/CONTRADICTION
	Neg_Head:	PersonX expresses PersonX's anger.		
ATOMIC2020 (Hwang et al., 2020)	**Head**:	PersonX takes a long walk.	**Hypothesis**:	PersonX takes a long walk.
	Relation:	HinderedBy	**Premise**:	It is 10 degrees outside..
	Tail:	It is 10 degrees outside.	**Label**:	CONTRADICTION
δ-NLI (Rudinger et al., 2020)	**Hypothesis**:	PersonX takes a long walk.	**Hypothesis**:	PersonX takes a long walk.
	Premise:	HinderedBy	**Premise**:	It is 10 degrees outside..
	Update:	It is 10 degrees outside.	**Label**:	CONTRADICTION
	Label:	Weakener		
CoreQuisite (Qasemi et al., 2021)	**Statement**:	A net is used for catching fish.	**Hypothesis**:	A net is used for catching fish.
	Precondition:	You are in a desert.	**Premise**:	You are in a desert.
	Label:	Disabling	**Label**:	CONTRADICTION

Table 10: Examples from target tasks in NLI format