PInKS: Preconditioned Commonsense Inference with Minimal Supervision

Ehsan Qasemi[1] and **Piyush Khanna**[2] and **Qiang Ning**[3] and **Muhao Chen**[1]
[1]University of Southern California [2]Delhi Technological University [3]Amazon
{qasemi,muhaoche}@usc.edu; piyushkhanna_bt2k17@dtu.ac.in;
qning@amazon.com

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 on benchmarks focused on reasoning with the preconditions of commonsense knowledge (up to 40% Macro-F1 scores). We further investigate *PInKS* through PAC-Bayesian informativeness analysis, precision measures, and ablation study.[1]

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 an agent to possess and understand different dimensions of commonsense knowledge (Woodward, 2011), e.g. physical, causal, social, etc. This ability can improve many knowledge-driven tasks such as 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 also seeks to benefit 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., 2022; 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 a common sense statement or chain of statements. For example, *PaCo* (Qasemi et al., 2022) approaches the task from the 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 from soft assumptions' perspective, i.e., *weakeners* and *strengtheners*, and justifies whether the *update* sentence *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 generally fall short of tackling these tasks.

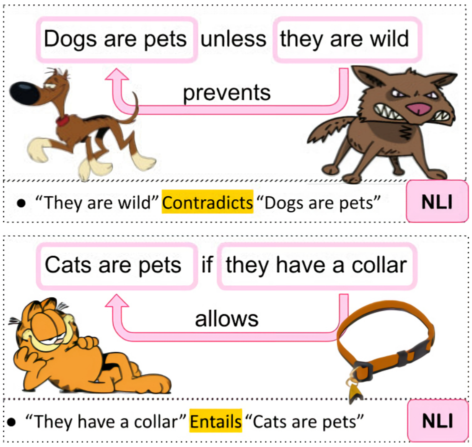
Dogs are pets
unless
they are wild
prevents
• "They are wild" Contradicts "Dogs are pets"
NLI
Cats are pets
if
they have a collar
allows
• "They have a collar" Entails "Cats are pets"
NLI

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

We identify two reasons for such shortcomings

[1] Code and data on https://github.com/luka-group/PInKS

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 320–336
November 20–23, 2022. ©2022 Association for Computational Linguistics

of LMs on reasoning with preconditions: 1) relying on expensive direct supervision and 2) the need for improved LMs to reason with such knowledge. First, current resources for preconditions of common sense are manually annotated. Although this yields high-quality direct supervision, it is costly and not scalable. Second, off-the-shelf LMs are trained on free-text 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 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 LMs with more targeted 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), *PaCo* (Qasemi et al., 2022) and DNLI (Rudinger et al., 2020)), we show that *PInKS* improves the performance of NLI models, up to 5% Macro-F1 without seeing any task-specific training data and up to 40% Macro-F1 after being incorporated into them (§4.1). In addition to the empirical results, using theoretical guarantees of informativeness measure in *PABI* (He et al., 2021), we show that the minimally supervised data of *PInKS* is as informative as fully supervised datasets (§4.2). Finally, to investigate the robustness of *PInKS* and effect of each component, we focus on the weak supervision part (§5). We perform ablation study of *PInKS* w.r.t. the linguistic patterns themselves, the recall value associated with linguistic patterns, and finally contribution of each section to overall quality and the final performance.

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; Davis and Marcus, 2015). The preconditions of common sense knowledge are eventualities that affect happening of a common sense statement (Hobbs, 2005). These preconditions can either *allow* or *prevent* the common sense statement in different degrees (Rudinger et al., 2020; Qasemi et al., 2022). For example, Qasemi et al. (2022) model the preconditions as *enabling* and *disabling* (hard preconditions), whereas Rudinger et al. (2020) model them as *strengthening* and *weakening*(soft preconditions). Beyond the definition of preconditions, the task of inference with preconditions is also defined differently among the literature. Some task definitions have strict constraints on the format of statement, e.g. two sentence format (Rudinger et al., 2020) or being human-related (Sap et al., 2019), whereas others do not (Do and Pavlick, 2021; Qasemi et al., 2022).

To unify the definitions in available literature, we define the preconditioned inference task as below:

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 the literature (for more details see appx. §G). First, similar to the definition by Rudinger et al. (2020), the update can have different levels of effect on the statement, from causal connection (hard) to material implication (soft). Second, similar to the one Qasemi et al. (2022), the statement can have any format.

3 Preconditioned Inference with Minimal Supervision

In *PInKS*, to overcome the challenges associated with inference with preconditions, we propose two

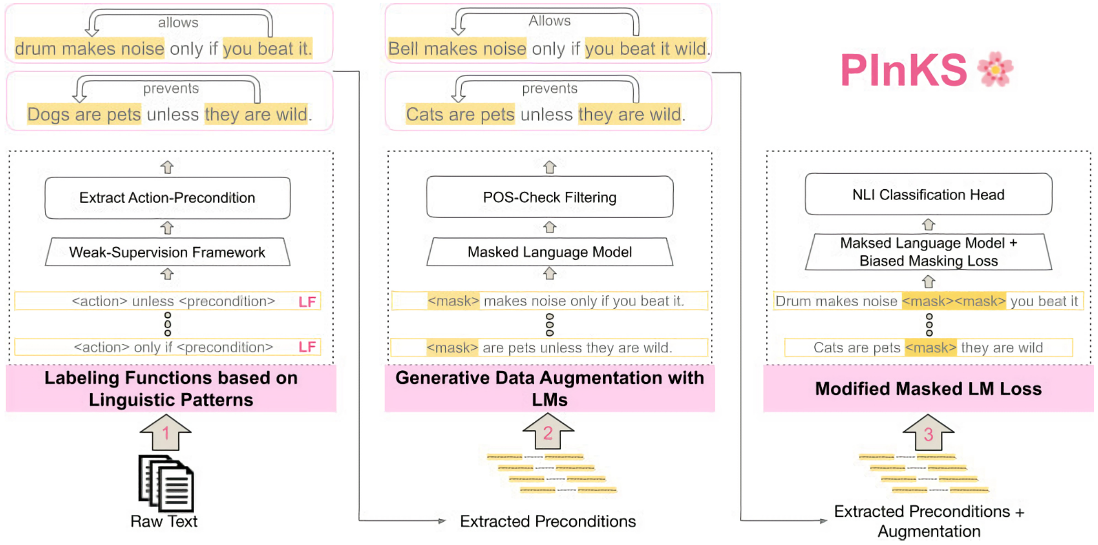
allows
drum makes noise only if you beat it.
prevents
Dogs are pets unless they are wild.
Allows
Bell makes noise only if you beat it wild.
prevents
Cats are pets unless they are wild.
PInKS
Extract Action-Precondition
Weak-Supervision Framework
<action> unless <precondition> LF
<action> only if <precondition> LF
Labeling Functions based on Linguistic Patterns
1
Raw Text
POS-Check Filtering
Masked Language Model
<mask> makes noise only if you beat it.
<mask> are pets unless they are wild.
Generative Data Augmentation with LMs
2
Extracted Preconditions
NLI Classification Head
Maksed Language Model + Biased Masking Loss
Drum makes noise <mask><mask> you beat it
Cats are pets <mask> they are wild
Modified Masked LM Loss
3
Extracted Preconditions + Augmentation

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

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 pretraining models on large amount of weak-supervised 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 an 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 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 *Abstaining* (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.

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.

Extracting Action-Precondition Pairs Once the sentence 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 could be sentences 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". Such co-occurring 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 precisions[2]. 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 scores of two annotators for each pattern/conjunction to get its precision score. This precision 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 precision in case of ambiguity. Further, we also set a minimum precision threshold (=0.7) to filter out the conjunctions having a low precision score (8 LFs), indicating low relevance to the task of reasoning with preconditions (see Appx. §A.1 for list of precision 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 (pivot-words) from the text and let the generative language model fill in the masks with appropriate alternatives.

After masking the pivot-word and filling in the mask using the 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 the LM's attention 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,

[2] The amounts of labeled instances (*non-abstaining*) for each labeling function are relevant

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. 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). We then theoretically justify the improvements by measuring the informativeness of weak supervision by *PInKS* using *PABI* (He et al., 2021) score and then experiment on the effect of precision (discussed in §3.1) on *PInKS* using *PABI* score (§4.2). Additional analysis on various training strategies of *PInKS* is also provided in Appx. §C.

4.1 Main Results

Comparing the capability for models to reason with preconditions across different tasks requires canonicalizing the inputs and outputs in such tasks be in the same format. We used natural language inference (NLI) as such a canonical format. *PaCo* (Qasemi et al., 2022) and δ-NLI (Rudinger et al., 2020) are already formulated as NLI and others can be converted easily using the groundwork laid by Qasemi et al. (2022). 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. 8.

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.[3] For more details on Snorkel and its setup details, please see Appendix A.2.

For each target task, we start from a pretrained NLI model (RoBERTa-Large-MNLI (Liu et al., 2019)), fine-tune it according to *PInKS* (as discussed in §3) and evaluate its performance on the test portion of the target dataset in two setups: zero-shot transfer learning without using the training data for the target task (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; *PInKS* is not used here). To create the test set, if the original data does not provide a split (e.g. *ATOMIC* and *Winoventi*), following Qasemi et al. (2022), 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 pretrained model 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 Task	Orig.	*PInKS*	Orig+*PInKS*
δ-NLI	83.4	60.3	**84.1**
PaCo	77.1	69.5	**79.4**
ANION	81.1	52.9	**81.2**
ATOMIC	43.2	**48.0**	**88.6**
Winoventi	51.1	**52.4**	**51.3**

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*). **Bold** values are cases where *PInKS* is improving supervised results.

Discussion Table 2 presents the evaluation results of this section. As illustrated, 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

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

additional data from target task to achieve better performance consistently across different aspects of preconditioned inference.

4.2 Informativeness Evaluation

He et al. (2021) proposed a unified PAC-Bayesian motivated informativeness measure, namely *PABI*, that correlates with the improvement provided by the incidental signals to indicate their effectiveness on a 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 explain how improvements from *PInKS* are theoretically justified. Here, we use the *PABI* score for cross-domain signal assuming the weak supervised data portion of *PInKS* (§3.1 and §3.2) as a indirect signal for a given target task. We use *PABI* measurements from two perspective. First, we examine how useful is the weak supervised data portion of *PInKS* for target tasks in comparison with fully-supervised data. And second, we examine how the precision of the linguistic patterns (discussed in §3.1) affects this usefulness.

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. §E. For the aforementioned first perspective, we only consider *PaCo* and δ-NLI as target tasks, as they are the two main learning resources specifically focused on preconditioned inference (as defined in Section 2), which is not the case for others. We measure the *PABI* of the weak supervised data portion of *PInKS* on the two target tasks, and compare it with the *PABI* of the fully-supervised data from §4.1. For the second perspective, we only focus on *PInKS* and consider *PaCo* as target task. We create different versions of the weak supervised data portion of *PInKS* with different levels of precision threshold (e.g. 0.0, 0.5) and compare their informativeness on *PaCo*. To limit the computation time, we only use $100K$ samples from the weak supervised data portion of *PInKS* in each threshold value, which is especially important in lower thresholds due to huge size of extracted patterns with low precision threshold.

Informativeness in Comparison with Direct Supervision: Tab. 3 summarizes the *PABI* informativeness measure in comparison with other datasets with respect to *PaCo* (Qasemi et al., 2022) and δ-NLI (Rudinger et al., 2020). To facilitate the comparison of *PABI* scores in Tab. 3, we have also reported the minimum achievable ("zero rate" classifier) and maximum achievable *PABI* scores. To clarify, to compute the maximum achievable *PABI* score, we consider the training subset of the target task as an indirect signal for the test subset. Here, we assume that the training subset is in practice the most informative indirect signal available for the test subset of any task. For the minimum achievable *PABI* score, we considered the error rate of the "zero rate" classifier (always classifies to the largest class) for computations of *PABI*.

	***PABI* on**		
Indir. Task	***PaCo***	**δ-NLI**	**Explanation**
PInKS	52.2	*66.7*	- Best on δ-NLI
δ-NLI	*52.3*	**85.5**	- Max achievable on δ-NLI - Best on *PaCo*
PaCo	**52.3**	31.3	- Max achievable on *PaCo*
ANION	34.1	13.9	
ATOMIC	20.9	17.4	
Winoventi	36.4	53.4	
Zero Rate	26.2	0.0	- Baseline

Table 3: *PABI* informativeness measures (x100) of *PInKS* and other target tasks w.r.t *PaCo* and δ-NLI. **Bold** values represent the maximum achievable *PABI* Score by considering train subset as an *indirect* signal for test subset of respective data. The highest *PABI* score, excluding the max achievable, is indicated in *italic*.

Our results show that although, *PInKS* is the top informative incidental signal in δ-NLI target task and second best in *PaCo* (less than 0.001 point of difference with the best signal). This *PABI* numbers are even more significant considering that *PInKS* is the only weak-supervision data which is automatically acquired, while others are acquired through sometimes multiple rounds of human annotations and verification.

Effect of Precision on Informativeness: Fig. 3 presents the *PABI* informativeness estimation on weak supervision data under different threshold levels of precision values, and compare them with the "zero rate" classifier (always predicting majority class). As illustrated, the informativeness show a significant drop in lower precision showcasing the importance of using high precision 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 again. This susceptibility on pattern precision

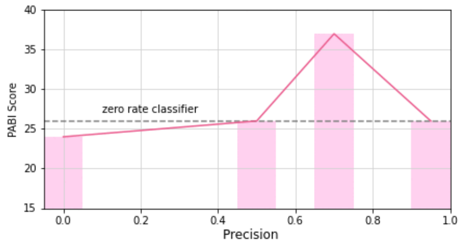

Figure 3: *PABI* informativeness measures of *PInKS* with different precision thresholds on *PaCo*.

can be mitigated with having more fine-grained patterns on larger corpora. We leave further analysis on precision of patterns to future work.

5 Analysis on Weak Supervision

In this section, we shift focus from external evaluation of *PInKS* on target tasks to analyze distinct technical component of *PInKS*. Here, through an ablation study, we try to answer four main questions to get more insight on the weak supervision provided by those components. First (Q1), how each labeling function (LF; §3.1) is contributing to the extracted preconditions? Second (Q2), what is the quality of the weak supervision data obtained from different ways of data acquisition? Third (Q3), how does generative data augmentation (§3.2) contribute to *PInKS*? And finally (Q4), how much does the precondition-aware masking (§3.3) affect the overall performance of *PInKS*?

(Q1) LF Analysis: To address the first question, we use statistics of the 6 top performing LFs (see Appx. §F for detailed results). These 6 top performing LFs generate more than 80% of data (Coverage) with the highest one generating 59% of data and lowest one generating 1%. Our results show that, in 0.14% of instances we have conflict among competing LFs with different labels and in 0.12% we have overlap among LFs with similar labels, which showcases the level of independence each LF has on individual samples.[4]

(Q2) Quality Control: To assess the quality of collected data, we used an expert annotator. The expert annotator is given a subset of the collected preconditions (preconditions-statement-label triplet) and asked to assign a binary label based on whether each the precondition is valid to its statement w.r.t the associated label. We then report the average quality score as a proxy for *precision* of data. We sampled 100 preconditions-statement-label triplets from three checkpoint in the pipeline: 1) extracted through linguistic patterns discussed in §3.1, 2) outcome of the generative augmentations discussed in §3.2, and 3) final data used in §3.3. Table Tab. 4 contains the average precision of the collected data, that shows the data has acceptable quality with minor variance in quality for different weak supervised steps in *PInKS*.

Checkpoint Name	Precision. %
Linguistic Patterns from §3.1	78
Generative Augmentation from §3.2	76
Final Data used in §3.3	76

Table 4: Precision of the sampled preconditions-statement-label triplets from three checkpoints in pipeline.

(Q3) Effectiveness of Generative Augmentation: The main effect of generative data augmentation (§3.2) is, among others, to acquire *PInKS* additional training samples labeled as *prevent* from pretrained LMs. When considering *PaCo* as target task, the *PInKS* that does not use this technique (no-augment-*PInKS*) sees a 4.14% absolute drop in Macro-F1 score. Upon further analysis of the two configurations, we observed that the no-augment-*PInKS* leans more toward the zero rate classifier (only predicting *allow* as the majority class) in comparison to the *PInKS*.

(Q4) Effectiveness of Biased Masking: We focus on *PaCo* 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 weak-supervision data obtained through *PInKS* to fine-tune the model and compare the results. Our results show that the Macro-F1 score for zero-shot transfer learning setup has a 1.09% absolute drop in Macro-F1 score, without the biased masking process.

6 Related Work

Reasoning with Preconditions 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 an *update* sentence *weakens* or *strengthens* a common sense hypothesis-premise pair. For example, given the premise "Two men and a dog are standing among rolling green hills.", the *update* "The men are studying a tour map" weakens

[4] Convectional inner-annotator agreement (IAA) methods hence are not applicable.

the hypothesis that “they are farmers”, whereas “The dog is a sheep dog” strengthens it. Similarly, *PaCo* (Qasemi et al., 2022) 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 *PaCo*, based on Shoham (1990) and Hobbs (2005), authors distinguish between two type of preconditions, causal connections (*hard*), and material implication (tends to cause; *soft*). 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. (2022), Do and Pavlick (2021) and Jiang et al. (2021). In these works, direct human supervision (crowdsourcing) is used to gather preconditions of commonsense knowledge. They all show the shortcomings of SOTA models on dealing 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; Singh et al., 2022), e.g. using heuristic rules. In addition, other methods such as re-purposing 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.

7 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 precision 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 (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 *PaCo* (Qasemi et al., 2022) they use ConceptNet as source of commonsense statements which multiple studies have shown its 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 may contain various types of bias. The labeling functions in *PInKS* are only limited to English language patterns, which may inject 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 D.

Limitations

The main limitation of this work are related to the choice of raw text corpora and the model for main results. From the raw text corpora perspective, we relied on Open Mind Common Sense (OMCS) (Singh et al., 2002) and ASCENT (Nguyen et al., 2021a) as two rich resource of commonsense knowledge. Future iterations of this work should include more fine-grained labeling functions to be applied to other large scale corpora that results in more diverse set of extracted preconditions.

The purpose of the experiments in this work is to show the effectiveness of *PInKS* in preconditioned inference without introducing any expensive (manually labeled) supervision. We chose RoBERTa-Large-MNLI (Liu et al., 2019) as a representative and strong model that has been widely applied to NLI tasks, including all those evaluated in this work. However, there are more models, e.g. unified-QA-11B for *PaCo* or DeBERTa for δ-NLI, that can be considered for each one of the target tasks. Of course achieving the SOTA with these much larger models requires a lot of computational resources, which is beyond the scope and bandwidth of this study. But, given more resources we would easily extend analysis to other models as well.

Acknowledgement

We like to thank Hangfeng He, for his insightful comments to using the *PABI* (He et al., 2021) measure in our paper. We also want to thank our anonymous reviewers whose comments/suggestions helped improve and clarify this paper.

This work is supported in part by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research, the National Science Foundation of United States Grant IIS 2105329, and a Cisco Research Award.

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 (Volume 2: Short Papers)*, pages 54–59, Jeju Island, Korea. Association for Computational Linguistics.

Stephen H. Bach, Bryan Dawei He, Alexander Ratner, and Christopher Ré. 2017. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 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 *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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, Prague, Czech Republic. Association for Computational Linguistics.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1790–1799, Online. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

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

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, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 65–74. ACM.

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, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

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 '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3023–3030. ACM.

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*.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. Association for Computational Linguistics.

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 *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3818–3828, Online. Association for Computational Linguistics.

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*.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

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

Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. 2021. skweak: Weak supervision made easy for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 337–346, Online. 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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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 Daniel 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, Suntec, Singapore. Association for Computational Linguistics.

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 *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5492–5500. IEEE Computer Society.

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

Ehsan Qasemi, Filip Ilievski, Muhao Chen, and Pedro Szekely. 2022. Paco: Preconditions attributed to commonsense knowledge. In *EMNLP-Findings*.

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 De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 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 *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

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.

Shikhar Singh, Ehsan Qasemi, and Muhao Chen. 2022. Viphy: Probing" visible" physical commonsense knowledge. *arXiv preprint arXiv:2209.07000*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Russell Stewart and Stefano Ermon. 2017. Label-free supervision of neural networks with physics and domain knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2576–2582. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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 '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

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. 5, summarizes the final patterns used in *PInKS*, coupled with their precision 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. 6 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. (2022), 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 translated 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. (2022), 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 dCONTRADICTION subset as it is clean of lexical hints. Then we convert their crowd-

Conjunctions	**Precision**	**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 5: 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 6: List of conjunctions used in modified masked loss function in section 3.3

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

Table 7: Filtered Labeling Functions Patterns and their associated polarity.

sourced *original head* 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 *PaCo* (Qasemi et al., 2022), 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. 8 summarizes the conversion process through examples from the original data and the NLI task derived from each.

To run all the experiments, we fine-tune the models on tuning data for maximum of 5 epochs with option for early stopping available upon 5 evaluation cycles with less than $1e - 3$ change on validation data. For optimizer, we use AdamW (Loshchilov and Hutter, 2019) with learning rate of 3e-6 and default hyperparamter for the rest.

C 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 monitor each task separately that increases interpretability of results.

However, in an alternative fine-tuning setup, one

Name	Original Data		Derived NLI	
Winoventi (Do and Pavlick, 2021)	**masked_prompt**: **target**: **incorrect**:	Margaret smelled her bottle of maple syrup and it was sweet. The syrup is {MASK}. edible malodorous	**Hypothesis**: **Premise**: **Label**:	Margaret smelled her bottle of maple syrup and it was sweet. The syrup is edible/malodorous ENTAILMENT/CONTRADICTION
ANION (Jiang et al., 2021)	**Orig_Head**: **Relation**: **Tail**: **Neg_Head**:	PersonX expresses PersonX's delight. xEffect Alice feel happy PersonX expresses PersonX's anger.	**Hypothesis**: **Premise**: **Label**:	Alice expresses Alice's delight/anger. feel happy. ENTAILMENT/CONTRADICTION
ATOMIC2020 (Hwang et al., 2020)	**Head**: **Relation**: **Tail**:	PersonX takes a long walk. HinderedBy It is 10 degrees outside.	**Hypothesis**: **Premise**: **Label**:	PersonX takes a long walk. It is 10 degrees outside.. CONTRADICTION
δ-NLI (Rudinger et al., 2020)	**Hypothesis**: **Premise**: **Update**: **Label**:	PersonX takes a long walk. HinderedBy It is 10 degrees outside. Weakener	**Hypothesis**: **Premise**: **Label**:	PersonX takes a long walk. It is 10 degrees outside.. CONTRADICTION
PaCo (Qasemi et al., 2022)	**Statement**: **Precondition**: **Label**:	A net is used for catching fish. You are in a desert. Disabling	**Hypothesis**: **Premise**: **Label**:	A net is used for catching fish. You are in a desert. CONTRADICTION

Table 8: Examples from target tasks in NLI format

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 weak-supervision 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 *PaCo*, δ-NLI and Winoventi as the target dataset.

Target Data	Orig+*PInKS* (Multi-Task)	Diff.
δ-NLI	72.1	-11.00
PaCo	77.3	+6.8
Winoventi	51.7	+0.7

Table 9: Macro-F1 (x100) results of *PInKS* on the target datasets using *multi-task* fine-tuning strategy and its difference with *curriculum* strategy.

Discussion Tab. 9 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 *PaCo* 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.

D Model Sizes and Run-times

All the experiments are conducted on a commodity workstation with an Intel Xeon Gold 5217 CPU and an NVIDIA RTX 8000 GPU. For all the fine-tuning results in Tab. 2, Tab. 3 we used "RoBERTa-Large-MNLI" with 356M tuneable parameters. To fine-tune the model in each experiment, we use Ray (Liaw et al., 2018) to handle hyperparameter tuning with 20 samples each. The hyperparameters that are being tuned fall into two main categories: 1) model hyperparameters such as "sequence length", "batch size", etc. and 2) data hyperparameters such as "precision threshold", "data size", etc.. The mean run-time for each sample 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 weak supervision data is 49hr that includes all three steps of data preprocessing, linguistic pattern matching, and generative data augmentation.

E 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 weak-supervision 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.

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

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,

Indir. Task	$\|L\|$	η_1	η_2^{ATMC}	η_2^{PaCo}	$\eta_2^{\delta-NLI}$	η^{ATMC}	η^{PaCo}	$\eta^{\delta-NLI}$	*PABI*^{ATMC}	*PABI*^{PaCo}	*PABI*$^{\delta-NLI}$
PInKS	2	0.04	0.11	0.21	0.16	0.076	0.202	0.129	0.782	0.523	0.667
δ-NLI	2	0.13	0.22	0.28	0.16	0.122	0.203	0.046	0.683	0.522	0.855
PaCo	2	0.03	0.10	0.22	0.33	0.074	0.202	0.318	0.786	0.523	0.313
ATOMIC	2	0.01	0.57	0.62	0.60	0.608	0.622	0.602	0.184	0.209	0.174
ANION	2	0.16	0.57	0.36	0.44	0.571	0.302	0.418	0.122	0.341	0.139
Winoventi	2	0.19	0.10	0.37	0.31	0.139	0.289	0.196	0.647	0.364	0.534

Table 10: Details of *PABI* metric computations in §4.2 according to Equation (1)

$\acute{\eta}_1$ 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 $\hat{S}(\pi_0, \tilde{\pi}_0)$ can be calculated as follows:

$$\sqrt{1 - \frac{\eta \ln(|\mathcal{L}| - 1) - \eta \ln \eta - (1 - \eta) \ln(1 - \eta))}{\ln|\mathcal{L}|}}$$

Tab. 10 contains the details associated computation of *PABI* score as reported in §4.2.

F Details on LFs in *PInKS*

Tab. 11 shows Coverage (fraction of instances assigned the non-abstain label by the labeling function), Overlaps (fraction of instances with at least two non-abstain labels), and Conflicts (fraction of instances with conflicting and non-abstain labels) on top performing LFs in *PInKS*.

LF name	**Cov. %**	**Over. %**	**Conf. %**
to understand	59.03	0.03	0.03
statement is	10.58	0.03	0.03
except	4.84	0.02	0.01
unless	4.79	0.04	0.04
in case	1.46	0.01	0.00
if not	1.00	0.01	0.01
Overall	81.69	0.14	0.12

Table 11: Coverage (fraction of raw corpus instances assigned the non-abstain label 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. Green and red color respectively represent LFs that assign *allow* and *prevent* labels.

G Details on Preconditioned Inference in the Literature

As mentioned in §2, existing literature does not have a consistent (unified) definitions from to aspects: 1) the definition of the preconditions, and 2) the definition of preconditioned inference.

First, existing literature define preconditions of common sense statements in different degrees of impact on the statement. For example, Qasemi et al. (2022) follows the notion of "causal complex" from Hobbs (2005), where for a common sense statement s preconditions of the statement $P_f(s)$ are defined as collection of eventualities (events or states) that results in s to happen. According to Qasemi et al. (2022), such eventualities can either *enable* ($p_f^+ \in P_f$) or *disable* ($p_f^- \in P_f$) the statement to happen. Also, Qasemi et al. (2022) uses Fikes and Nilsson (1971) to define *disable* as *enabl*ing the negation of the statement. On other hand, Rudinger et al. (2020) defines *strengthener* as updates that a human would find them to increase likelihood of a hypothesis, and the *weakener* as the one that humans would find them to decrease it. Here, the focus on human's opinion is stemmed from definition of common sense. In this work, given the focus on noisy labels derived from weak-supervision, we adopted the more relaxed definition from Rudinger et al. (2020) for preconditions of common sense statements.

Second, there is also inconsistencies in the definition of reasoning with the preconditions or preconditioned inference. Rudinger et al. (2020) has a strict structure. It defines the task w.r.t to effect of precondition on the relation of two sentences: hypothesis and premise; where a model has to find the type of the precondition based on whether it *strengthens* or *weakens* the relation between the two sentences. Differently, Qasemi et al. (2022) has a relaxed definition in which the model is to decide if the precondition either enables or *disables* the statement. Here the statement can have any format. Do and Pavlick (2021), Hwang et al. (2020), and Jiang et al. (2021), on the other hand, define only a generative task to evaluate the models. In this work, again we adopted the more relaxed definition from Qasemi et al. (2022) that imposes less constraint on weak-supervised data.