Evaluate Confidence Instead of Perplexity for Unsupervised Commonsense Reasoning

Anonymous ACL submission

Abstract

In this paper, we present a novel approach to unsupervised commonsense reasoning that outperforms conventional perplexity evaluation. Specifically, we propose the use of non-replacement confidence (NRC), which is evaluated by a pre-trained token corruption discriminator. We show that NRC is a more consistent metric for commonsense reasoning, as it allows for equal synonym positiveness and negative sample learning. Our experiments using the ELECTRA discriminator demonstrate that NRC significantly outperforms perplexity on both tuple and sentence-level commonsense knowledge databases. Moreover, we show that NRC sets a new unsupervised state-of-the-art (SOTA) on seven commonsense question answering tasks, outperforming even complex reasoning systems. In supervised learning, we find that NRC is the most successful metric for applying pre-trained knowledge on annotated data for inference. In fact, without negative samples, NRC achieves between 82.8% and 90.0% of the performance of supervised methods, significantly outperforming other metrics under weaker supervision. To further improve the performance of NRC, we propose a new scenario in which the discriminator is first pre-trained on positive samples and then the NRC evaluation of negative samples is incorporated to tune the confidence. This approach significantly outperforms conventional fine-tuning by an average of 2.0 accuracy points. In summary, our research indicates that NRC is a superior metric compared to perplexity when it comes to learning commonsense knowledge under various supervision settings.[1]

1 Introduction

Commonsense reasoning is the underlying basis of machines for human-like natural language understanding. Commonsense knowledge endows natural language processing (NLP) systems with the awareness of implicit background for how human inference deals with the physical world. External commonsense knowledge created by humans has been successfully applied to refine NLP systems like dialogue (Zhou et al., 2021) and generation (Chakrabarty et al., 2021).

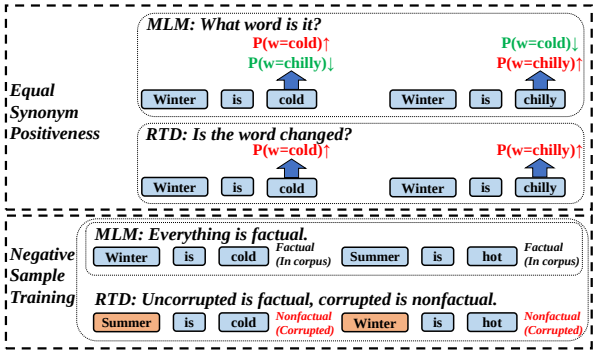

Figure 1: The differences between the generative and discriminative evaluation of factual consistency.

For unsupervised commonsense reasoning, perplexity has long been applied to estimate how a piece of expression is consistent with common facts. Originating from the statistical language model, text with higher perplexity is estimated to have less probability to appear in natural language, and thus less consistent with facts. This idea is further strengthened by the emergence of large pre-train neural language models (PLMs), which introduce deep Transformer encoders and more advanced training objectives, like masked language modeling (MLM) (Devlin et al., 2019).

While PLM-based perplexity shows potential for commonsense understanding and has become the basic component of many more advanced unsupervised methods (Shwartz et al., 2020; Bosselut et al., 2021; Niu et al., 2021), the diversity of PLMs also suggests perplexity is no longer the only way to estimate how texts are consistent with facts. Besides generative PLMs that predict unseen words, ELECTRA (Clark et al., 2020) uses a discriminator

[1] Our code is released at github.com/KomeijiForce/ELECTRA-NRC

pre-trained to detect token corruption in context. This training objective is named replaced token detection (RTD), which takes texts corrupted by MLM as input and predicts whether each token is corrupted or not.

The RTD training objective is more consistent with commonsense reasoning for two reasons as shown in Figure 1. **1. RTD discriminator allows positive synonyms to share equal positiveness. (Equal Synonym Positiveness)** For generative probability, all candidates in the dictionary share a probability distribution sum up to 1. Consequently, if a word, unfortunately, has many synonyms in the dictionary, most of its generative probability will be taken by those synonyms. **2. RTD discriminator is pre-trained on negative samples. (Negative Sample Learning)** In commonsense reasoning, models always face non-sense expressions that distract the reasoning. However, pre-trained generators only learn positive samples and thus become inapt at evaluating unseen negative samples. In contrast, the RTD discriminator naturally classifies corrupted tokens that turn texts into factually inconsistent.

To validate and probe the advantages of the RTD discriminator, we derive a metric called non-replacement confidence (NRC) from its training objective. We apply NRC for supervised and unsupervised commonsense reasoning and compare its performance with conventional perplexity-based inference. NRC outperforms the best perplexity-based methods on tuple and sentence-level commonsense knowledge databases, together with 7 commonsense question answering tasks. NRC even outperforms previous complex systems to set the new unsupervised state-of-the-art (SOTA) on all question answering datasets. To validate claimed advantages, we use a synonym replacement attack to show the ELECTRA discriminator how to react to words with synonyms. NRC is more robust against synonym replacement attacks than PPL. We also explore how negative samples created by the ELECTRA generator affect knowledge capturing. In comparison with PPL, NRC shows a higher capability to learn from knowledge sources and mask rate plays a critical role in the qualifying of captured knowledge.

The advantages of the ELECTRA discriminator are notable, and they motivate us to delve deeper into the potential of NRC for supervised learning. Our experimental results indicate that by pre-training the ELECTRA discriminator solely on positive samples from the training dataset, we can achieve a performance ranging from 82.8% to 90.0% compared to supervised methods trained on the entire dataset. Then, we use the NRC differences between positive and negative samples to tune those discriminators, their performance outperforms direct fine-tuning by 2.0 accuracy scores on average. As the backbone model in the current SOTA model (Xu et al., 2021, 2022), DeBERTaV3 (He et al., 2021), is also an RTD-based PLM, our discovery shows the potential to further boost the supervised SOTA[2].

Our contributions are summarized as follows:

- We suggest using a pre-trained discriminator objective as an alternative to pre-trained generators for commonsense reasoning. Our study shows that the NRC metric, derived from the RTD objective, is more appropriate for the task as it supports equal synonym positiveness and negative sample learning.

- To explore unsupervised commonsense reasoning, we conducted experiments on NRC using tuple and sentence-level commonsense knowledge databases and evaluated it on 7 commonsense question answering tasks. Our results show that NRC outperforms perplexity and even outperforms other complex systems, achieving the new SOTA on all datasets.

- In supervised commonsense reasoning, we discovered that pre-training the ELECTRA discriminator on only positive samples resulted in a performance of 82.8% to 90.0%, compared to fully supervised methods. When incorporating negative samples, NRC showed an average increase in accuracy of 2.0 points compared to direct fine-tuning, which can be attributed to the learning of differences among samples.

2 Background

2.1 Commonsense Knowledge

Commonsense knowledge, also known as background knowledge, is the underlying basis of logic in the inference of humans. As commonsense knowledge is rarely expressed in textual contents

[2]We cannot apply confidence tuning for DeBERTaV3 because its generator has not been released yet.

(Gordon and Durme, 2013), many datasets (Bollacker et al., 2008; Nickel et al., 2011; Yang et al., 2015; Li et al., 2016) have been handcrafted to train NLP systems and endow them with the ability to make a physical world-based inference.

Following the storage system in databases, commonsense knowledge is generally formalized as a tuple (LT, RT, REL), e.g. ConceptNet (Speer and Havasi, 2012; Speer et al., 2017). Here, LT, RT, REL respectively refer to the left term, the right term, and the relationship between two terms. While tuples are efficient for storage, they are incompetent to represent relationships with more than 2 terms. Thus, Wang et al. create a sentence-level commonsense dataset, which validates the integrity of commonsense in a real context. This dataset also includes commonsense explanations for facts, which further expands the coverage of knowledge. Other commonsense knowledge bases are also introduced in recent years like ATOMIC (Sap et al., 2019; Hwang et al., 2021) for If-Then relationships. TransOMCS (Zhang et al., 2020) retrieve ConceptNet-like tuples from syntactic structures.

2.2 Commonsense Reasoning with PLMs

Large-scale pre-trained language models like BERT (Devlin et al., 2019) have drawn the most attention from the NLP community since their introduction. PLMs show their potential to significantly boost performance on NLP tasks across fields. Since PLMs have been trained on a large-scale corpus to learn interdependency between components, mining from PLMs for commonsense knowledge becomes a new method to create knowledge databases (Petroni et al., 2019; Alghanmi et al., 2021; Kassner et al., 2021). LAMA (Petroni et al., 2019) makes the first try to gather knowledge from PLMs by generative prompts. Later works follow this process to provide partial information in the commonsense knowledge tuple and require PLMs to complete the rest of the tuple.

The commonsense knowledge and understanding of PLMs inspire researchers to directly apply PLMs for downstream inference without supervised fine-tuning. Commonsense question answering (Roemmele et al., 2011; Zellers et al., 2018; Talmor et al., 2019, 2022; Kocijan et al., 2020) is commonly used to test the unsupervised inference ability of PLMs. Similar to commonsense reasoning, prompts are applied to transform the question-answer pair into a syntactically plausible sentence. PLM-based perplexity is calculated for those transformed sentences and the sentence with the lowest perplexity is used to select the corresponding question-answer pair (Trinh and Le, 2018; Bosselut et al., 2021; Tamborrino et al., 2020). Besides direct reasoning on answer candidates, researchers have also tried to sample extra candidates from generators and use pre-trained semantic similarity evaluator for answer selection. (Shwartz et al., 2020; Niu et al., 2021; Bosselut et al., 2021)

Current mainstream PLMs, BERT or GPT2, apply the conventional perplexity metric to use the probability of generating components based on the context. This will incorporate surface forms like word frequency as perturbance to the inference. Answer-Level Calibration (Kumar, 2022) models context-independent biases in terms of the probability of a choice without the associated context, and removes them using an unsupervised estimate of similarity with the full context for question answering tasks. Pointwise Mutual Information (Holtzman et al., 2021) factors out the probability of specific surface forms and introduce scoring-by-premise to measure the probability of the premise given the hypothesis. Based on the nature of commonsense reasoning, we propose a pre-trained discriminator, like ELECTRA, to be an alternative for better performance.

3 PLM-based Metric

3.1 Perplexity

Casual Language Model GPT2 (Radford et al., 2019) is a PLM pre-trained for text generation, which can also be applied for inference based on the perplexity of selection candidates. The training objective, CLM, is optimized based on context-based next-word prediction.

$$\mathcal{L} \triangleq \text{CELoss}(\text{PLM}_\theta(w_{1:i-1}), \text{One-hot}(w_i))$$

where CELoss is the cross-entropy loss, and One-hot refers to the one-hot encoding. θ, w respectively refer to PLM parameters and words. The inference procedure also takes next-word prediction for perplexity (PPL) calculation.

$$PPL = \frac{1}{n}\sum_{i=1}^{n}(-\log(p(w_i|\text{PLM}_\theta, w_{1:i-1})))$$

where n is the length of the sentence. GPT2 calculates PPL by scoring answer choices and selecting a candidate with the lowest perplexity.

Masked Language Model MLM is the training objective for most bidirectional PLMs like BERT and RoBERTa (Liu et al., 2019). MLM is similar to CLM as it also uses word retrieval as the training objective but leverages the bidirectional context for the prediction. During the inference, the likelihood of each word is calculated by a mask-and-predict procedure.

3.2 Replaced Token Detection

RTD differs from the word retrieval-targeted training procedure above as it sets binary classification as the objective. The PLM involves a discriminator which discerns replaced words in the sentence by an MLM-based generator.

$$\mathcal{L} \triangleq \text{BCELoss}([\text{PLM}_\theta(w_{1:n})]_i, f_B(w_i))$$

where f_B is a Boolean function that returns whether w_i is corrupted by the replacement or not.

We then derive the Non-Replacement Confidence metric from the training objective.

$$\textit{NRC} = \frac{1}{n}\sum_{i=1}^{n}(-\log([\text{PLM}_\theta(w_{1:n})]_i))$$

3.3 Metric Comparison

PPL and NRC are both calculated based on negative log probability. While PPL evaluates the likelihood of a sentence, NRC reflects the confidence of contextual integrity. Thus, lower PPL and higher NRC on legal language indicate more human-like choices.

Commonsense reasoning expects to understand the underlying interdependency between abstract concepts rather than their surface forms. Thus, evaluating confidence in the piece of commonsense knowledge should include not only words in the original sentence but their contextual synonyms as well.

$$p_{CS}(w_{1:n}) = \sum_{w \in \text{syn}(w_i)} p(C_i)p(w|C_i)$$

where p_{CS} is the **c**ommon**s**ense-targeted confidence. $C_i = w_{1:i-1;i+1:n}$ refers to the context for w_i and syn returns the contextual synonyms of w_i. As $w_i \in \text{syn}(w_i)$, $p_{CS}(w_{1:n}) > p(w_{1:n}) = PPL$ when the number of synonym candidates is more than 1, indicating that perplexity always underestimates the commonsense-targeted confidence. The underestimation becomes more severe when w_i is a low-frequency word. Furthermore, as $\sum_{w \in dict}(p(w)) = 1$ ($dict$ is the whole dictionary for token selection), the correlation between confidence on synonym candidates is -1. This indicates $-\frac{\partial \mathcal{L}}{\partial p(w=w_i)} > 0$ while $-\frac{\partial \mathcal{L}}{\partial p(w=w_j)} < 0, w_j \in \text{syn}(w_i)$ during the gradient updating. Thus, $p(w = w_j), w_j \in \text{syn}(w_i)$ is decreased by $-\frac{\partial \mathcal{L}}{\partial p(w=w_j)}$, which is contrary to the nature that a word supports the appearance probability of its synonyms.

In contrast to PPL, NRC evaluates the confidence of each candidate individually, without requiring them to share the same distribution. This means that there is no bias towards high-frequency words or underestimation due to underlying synonym candidates. Additionally, PLMs project contextually similar components to near positions in the latent space (Devlin et al., 2019), which changes the correlation between synonym candidates to positive. As a result, NRC provides a more accurate evaluation of the confidence of commonsense knowledge in a given context.

Another advantage of NRC comes from the pre-training process of ELECTRA. During the pre-training, the ELECTRA generator corrupts tokens which makes the input to the discriminator similar to the negative samples during testing. In contrast, pre-trained generators only take legal texts as the input and thus discern nonfactual expressions because they have not seen them during pre-training.

We use experiments to further explore the correctness of our theoretical analysis. In § 4.1, 4.2, we provide a rough view of how much better are NRC than PPL. In § 5.1, we use a synonym replacement attack to explore how NRC and PPL react to words with different frequencies. In § 5.2, 5.3, we explore how the further pre-training on in-domain and out-of-domain positive question-answer pairs affects the inference performance.

We also compare the time complexity of different metrics. Our NRC is $O(1)$ (counting the number of PLM forwarding) as efficient as the CLM-based inference since the discriminator does not use mask tokens to calculate the metric, which limits the efficiency of MLM-based inference to $O(n)$.

Metric	ConceptNet	SemEval_A	SemEval_B
$\text{PPL}_{\text{GPT2-XL}}$	65.4	78.1	58.1
$\text{PPL}_{\text{GPT2-M}}$	49.6	50.1	40.3
PPL_{BERT}	66.2	76.2	54.4
$\text{PPL}_{\text{RoBERTa}}$	69.9	79.9	62.4
NRC	**71.2**	**80.5**	**64.3**

Table 1: Experiment results on tuple and sentence-level commonsense reasoning. **Bold:** The best performance on the dataset. Underline: The result is significantly better than the second-best result. ($\alpha = 0.01$)

4 Unsupervised Inference

To mitigate the unfair comparison caused by the parameter scales, this paper compares large models with the same number of layers and hidden sizes, namely $\textbf{BERT}_{\text{Large}}$, $\textbf{RoBERTa}_{\text{Large}}$, $\textbf{GPT2}_{\text{Medium}}$ and $\textbf{ELECTRA}_{\text{Large}}$ (24-layer, 1024-hidden size). We also include $\textbf{GPT2}_{\text{XLarge}}$ (48-layer, 1600-hidden size) for further comparison.

4.1 Commonsense Probing

Tuple-level Probing ConceptNet[3] uses deep neural networks to retrieve commonsense candidates from corpus, which are validated by human annotators. We use the test dataset from (Li et al., 2016) which requires models to discern between true commonsense tuples and adversarial fake ones.

To create prompts for tuples in the test dataset that can be represented in natural language, we followed the methodology outlined in LAMA (Petroni et al., 2019). The full list of prompts can be found in Appendix C. Next, we utilize PLM-based metrics to distinguish between different prompts and evaluated the accuracy of our approach. Specifically, we evaluate PPL and NRC on templatized tuples and select the top 50% of tuples as positive results.

The classification accuracy is presented in Table 1, NRC significantly outperforms both CLM and MLM-based PPL on commonsense tuple reasoning. As transformed tuple relationships are simple in syntactic structures, we attributed the discriminating ability to the understanding of commonsense, which supports the superiority of NRC in commonsense validation.

[3]https://conceptnet.io/

Sentence-level Probing SemEval2020[4] collects natural language statements related to commonsense expression. We experiment with two reasoning subtasks. **A:** Select a statement that is against the commonsense. **B:** Select a reason to support the selection in **A**. We continue evaluating and selecting statements and explanations according to different metrics.

Table 1 also shows experiment results on sentence-level commonsense reasoning using various metrics. NRC performs significantly better than PPL on both differentiating and explanation, indicating its superior capability for sentence-level commonsense evaluation. $\text{PPL}_{\text{RoBERTa}}$ is competitive for differentiating but lags behind NRC in explanation since it requires a more complex inference ability. Overall, the comparison supports NRC as a more competent metric for commonsense reasoning.

4.2 Commonsense Question Answering

Baselines Besides perplexity, we also include previous trials for knowledge extraction (ALC (Kumar, 2022), PMI_{DC} (Holtzman et al., 2021)) and more complex baselines (Self-Talk (Shwartz et al., 2020), CGA (Bosselut et al., 2021), SEQA (Niu et al., 2021), ArT (Wang and Zhao, 2022)). Since SEQA used annotated NLI data for training, we follow (Wang and Zhao, 2022) to report the result from the original paper as SEQA_{NLI} and w/o NLI data version from (Wang and Zhao, 2022) as $\text{SEQA}_{\text{Orig.}}$ More detailed descriptions of baselines can be found in Appendix B.

Datasets We conduct experiments on a wide range of datasets to reach a more general conclusion. We test different methods on 7 datasets, including 2 phrase selection datasets {CommonsenseQA (CSQA), AI2 Reasoning Challenge (ARC)}, 2 entailment datasets {Choice of Plausible Alternatives (COPA), Situations With Adversarial Generations (SWAG)}, and 3 context-based datasets {StoryClozeTest (SCT), SocialIQA (SQA), CosmosQA (CQA)}. Specific descriptions and instances from those datasets are presented in Appendix A. For Phrase Selection, COPA, and SQA, we transform question-answer pairs by templates (Ma et al., 2021) to achieve comparable baselines with previous works.

[4]https://github.com/wangcunxiang/SemEval2020-Task4-Commonsense-Validation-and-Explanation

Method	Trg.	Phrase Selection			Entailment		Context-Based			Avg.
		CSQA	ARC_E	ARC_C	COPA	Swag	SCT	SQA	CQA	
Self-Talk	-	32.4	-	-	68.6	-	70.4	47.5	36.1	-
CGA	-	-	-	-	72.2	-	71.5	45.4	42.2	-
$SEQA_{Orig.}$	-	-	-	-	54.4	-	54.9	36.6	-	-
$SEQA_{NLI}$	-	-	-	-	79.4*	-	83.2*	47.5*	56.1*	-
ArT	-	-	-	-	69.8	-	71.6	47.3	-	-
ALC	-	49.7	-	-	81.6	-	-	45.1	-	-
PMI_{DC}	-	50.3	51.5	33.0	77.0	-	-	-	-	-
$PPL_{GPT2\text{-}XL}$	A	42.6	50.8	28.8	73.6	65.3	70.6	45.5	35.5	51.6
$PPL_{GPT2\text{-}M}$	A	38.5	44.4	24.9	68.4	59.7	54.0	44.3	27.0	45.0
PPL_{BERT}	Q	40.6	37.2	26.7	64.2	44.5	63.5	39.6	32.9	43.7
	A	28.0	37.1	22.7	61.2	63.4	58.2	40.4	30.7	42.7
	QA	32.8	36.8	23.7	64.2	64.1	61.2	38.5	29.6	43.9
$PPL_{RoBERTa}$	Q	49.3	40.5	35.6	70.6	48.1	61.5	39.7	38.6	48.0
	A	39.8	44.2	27.1	68.4	71.0	67.3	45.5	36.1	49.9
	QA	49.0	45.5	31.8	75.2	74.5	71.7	46.2	36.5	53.8
NRC	Q	51.2	46.8	38.6	**82.6**	24.5	65.0	40.6	41.2	48.8
	A	45.0	47.9	37.1	71.2	**77.4**	74.7	46.1	41.9	55.2
	QA	**54.1**	**52.1**	**39.8**	78.4	75.4	**77.1**	**47.7**	**44.3**	**58.6**

Table 2: Results on Unsupervised Commonsense Reasoning. Underline: A significant improvement compared to the best perplexity-based method. **Bold:** The best performance among unsupervised models ($SEQA_{NLI}$ is excluded because it requires NLI data for training). *: This result is obtained by an NLI-based zero-shot inference.

Phrase Selection We have adopted the approach of previous studies (Brown et al., 2020; Shwartz et al., 2020; Niu et al., 2021) to evaluate different targeted components (Question (Q), Answer (A), Question+Answer (QA)) for inference. The selection results are presented in Table 2. Our findings show that NRC outperforms PPL based on PLM with the same scale by a significant margin (4.8, 6.6, 4.2 accuracy score), which highlights NRC's superiority in using commonsense for inference. For the easy part of ARC (ARC_E), large-scale models like $GPT2_{XL}$ appear to be able to compensate for bias in the metric. However, as the questions become more challenging in ARC_C, the gap between NRC and PPL widens to 6.8 accuracy scores, underscoring the inherent differences between NRC and PPL in commonsense reasoning ability. NRC also outperforms Self-Talk, ALC, and PMI_{DC} to set the new SOTA.

Entailment NRC has once again demonstrated its superiority over PPL with an impressive performance on COPA and Swag (7.4 and 2.9, respectively). This has been validated by the large Swag dataset, confirming NRC's excellence in commonsense understanding. Furthermore, NRC's exceptional performance outshines complex systems for COPA, pushing the boundaries of the state-of-the-art. However, it is worth noting that the question part of Swag may not be very useful for NRC, as these questions are not answer-dependent from the perspective of ELECTRA. Instead, NRC prefers to use the answer portion of the dataset for inference. Nevertheless, when evaluating the entire question-answer pair, NRC consistently outperforms PPL.

Context-based The NRC model has demonstrated superior performance compared to PPL-based models, particularly on a large scale like $GPT2_{XLarge}$. This gap widens on datasets with longer context, such as SCT and CQA, indicating NRC's ability to comprehend complex contexts and the interdependencies between terms. While SEQA appears to hold the current SOTA on context-based selection, recent research (Wang and Zhao, 2022) has shown that its reasoning abilities stem from pre-training on NLI tasks. When NLI pre-training is removed, SEQA's performance drops sharply. Therefore, NRC currently sets the new state-of-the-art on all seven commonsense question answering tasks, surpassing several complex reasoning systems and the large GPT2-XL model.

Method	Accuracy (↑)	Affected Ratio (↓)
$\text{PPL}_{\text{GPT2-M}}$	47.2	30.4
PPL_{BERT}	58.0	30.2
$\text{PPL}_{\text{RoBERTa}}$	64.4	25.6
NRC	**72.4**	**22.4**

Table 3: Affect of synonym replacement on different inference methods. **Accuracy** is the ratio of correct selections after the replacement. **Affect Ratio** refers to the ratio of previous correct selections that are turned into faults by the replacement.

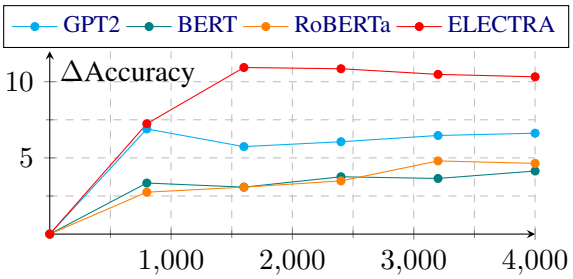

Figure 2: Improvements gained from further pre-training on question-answer pairs.

5 Further Analysis

5.1 Equal Synonym Positiveness

We validate the advantage of NRC-based inference when facing words with multiple synonyms by testing the accuracy of answer selection after synonym replacement. For implementation, we sample synonyms from Wordnet in NLTK to corrupt 10% words in each question and answer text of the COPA dataset.

The results of our experiments are presented in Table 3. Our NRC retains the highest performance compared to other metrics and still keeps a large margin. Also, NRC is the least likely to be affected by the replacement. Thus, the superiority of NRC over PPL facing synonyms is verified.

5.2 Negative Sample Learning

We explore the effect of negative sample learning in this section. We follow the idea in ELECTRA to compare the model performances trained by different steps. Thus, we further pre-train PLMs on question-answer pairs formalized as *Q[SEP]A*. These question-answer pairs can be seen as a new knowledge source. We test on CSQA to investigate how different pre-training methods obtain knowledge.

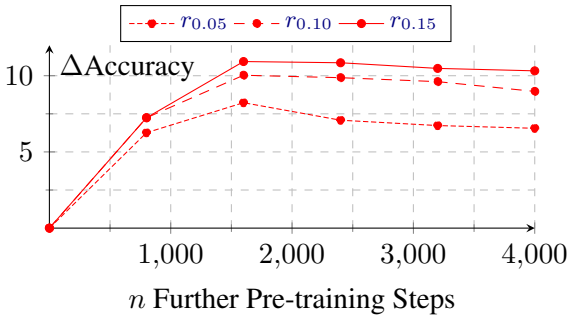

Figure 3: Improvements gained from further pre-training of ELECTRA with different mask rates r.

We use a question generator[5] to generate questions about noun chunks in the Wikipedia corpus. We train each PLM for 4000 steps with batch size 32 and report the best performance among models saved for each 800 steps. In Figure 2, we demonstrate the reasoning performance of language models on each step. We can observe that pre-trained generators obtain the understanding of question-answer in 800 steps but fail to further improve their performances. In comparison, the pre-trained discriminator steadily obtains knowledge in 3200 steps to capture the knowledge from question-answer pairs. It is also worth mentioning that GPT2 achieves high improvement in 800 but then begins to perform worse. This indicates GPT2 can quickly capture the question-answer syntactic structure but is poor at obtaining knowledge. We attribute this to the uni-direction of GPT2, which limits its ability in building connections between components.

We further explore the effect of negative samples in Figure 3 by changing the mask rate during training. $r_{0.15}$ is the mask rate in the initial training configuration. The results show that lowering mask rates lead to poorer performances, which verifies the benefit of negative samples in pre-training for commonsense knowledge capturing.

5.3 Supervised Confidence Tuning

Inspired by the prominent performance of RTD-based further pre-training on out-of-domain data, we propose confidence tuning for supervised learning. Based on the negative augmentation property of the pre-training process, we are first interested in training a supervised model without negative samples, which cost the energy of annotators to create adversely.

[5] https://huggingface.co/mrm8488/t5-base-finetuned-question-generation-ap

Method	CSQA	ARC_E	ARC_C
PPL_{GPT2-M}	55.0	48.2	27.7
PPL_{BERT}	50.5	43.3	27.9
$PPL_{RoBERTa}$	59.1	44.2	32.6
NRC	**73.0**	**59.9**	**45.1**

Table 4: Performance of PLMs further pre-trained on positive in-domain samples.

Method	CSQA	ARC_E	ARC_C
FT w/ Neg.	81.1	71.5	54.5
CT w/o Neg.	73.0	59.9	45.1
CT w/ Neg.	**83.2**	**73.6**	**56.5**

Table 5: Comparison between fine-tuning and confidence tuning.

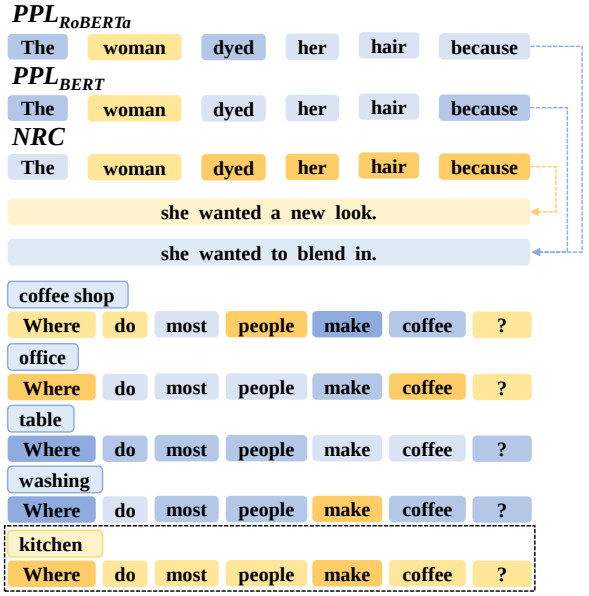

Figure 4: A case study on the reasoning of NRC.

Training w/o Negative Samples We run the same pre-training process on the CSQA training dataset and test the performance of different generative PLMs for comparison. The experiment results are presented in Table 4. While all PLMs benefit from the application of data created by humans, the gap between the performance of generative and discriminative PLMs remains large. The generative models are still inept at applying knowledge learned during pre-training even using the human-annotated datasets because of the unseen negative answers in the test dataset.

In comparison, the RTD training objective helps the ELECTRA discriminator to capture the commonsense knowledge embodied in the training data as the corrupted tokens turn the question-answer pairs to nonfactual, which are similar to negative samples in the test dataset. Table 5 further compares our training result to fine-tuning w/o negative samples. Remarkably, RTD pre-training reaches 82.8% (ARC_C)$\sim$ 90.0% (CSQA) of the supervised performance among the 3 datasets.

Training w/ Negative Samples Based on the further pre-trained discriminator, we tune it on negative samples with the application of cross entropy loss on the probability distribution $P(A_i|Q) = \frac{\exp(\text{NRC}(Q[\text{SEP}]A_i))}{\sum_j \exp(\text{NRC}(Q[\text{SEP}]A_j))}$. Table 5 shows the performance of our multi-stage learning method. Here we also fine-tune the ELECTRA discriminator already further pre-trained on positive question-answer pairs to make the comparison fair. Learning the confidence differences between positive and negative samples boost the supervised learning results by 2.1, 2.1, 2.0 accuracy scores on the three datasets. Our confidence tuning also achieves improvement in cross-dataset transfer learning, especially in transferring the knowledge learned on the challenging ARC_C dataset.

5.4 Case Study

We provide cases to specify the observation in statistics. The first case on COPA shows the limited understanding of PPL on the low-frequency phrase *dyed her hair*. NRC instead successfully leverages the semantics of the phrase to select the right answer. The second case on CommonsenseQA shows NRC to infer based on *Where* and *make coffee* and selects the answer supported by both key phrases, verifying its reasoning to be highly interpretable.

6 Conclusion

We propose a novel method to apply the training objective of a pre-trained discriminator rather than a generator for commonsense reasoning. The metric Non-Replacement Confidence, derived from the replaced token detection learning objective, better estimates textual consistency to facts by allowing equal synonym positiveness and negative sample learning. Unsupervised experiments verify the advantages of NRC over perplexity on commonsense knowledge probing and question answering. We further utilize the discovery in unsupervised learning and apply confidence tuning to supervised learning, which reaches desirable performance on learning with or without negative labels.

Limitation

Firstly, although NRC has shown impressive performance in unsupervised learning, the results still lag behind those achieved by supervised methods. NRC achieves between 82.8% and 90.0% of the performance of supervised methods under weaker supervision, indicating that there is still room for improvement to close the performance gap. Secondly, our study utilizes the existing ELECTRA models to evaluate the NRC metric, and the limited scale of these models restricts our ability to thoroughly compare the performance of NRC across a wider range of model scales. As larger and more powerful language models continue to be developed, it will be crucial to assess the performance of NRC with these models to better understand the metric's efficacy in various contexts and examine its scalability when applied to different model architectures.

References

Israa Alghanmi, Luis Espinosa Anke, and Steven Schockaert. 2021. Probing pre-trained language models for disease knowledge. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3023–3033. Association for Computational Linguistics.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4923–4931. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. 2021. Implicit premise generation with discourse-aware commonsense knowledge models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6247–6252. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, pages 25–30. ACM.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7038–7051. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: investigating knowledge

in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3250–3258. Association for Computational Linguistics.

Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. 2020. A review of winograd schema challenge datasets and approaches. *CoRR*, abs/2004.13831.

Sawan Kumar. 2022. Answer-level calibration for free-form multiple choice question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 665–679. Association for Computational Linguistics.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer 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. *CoRR*, abs/1907.11692.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 809–816. Omnipress.

Yilin Niu, Fei Huang, Jiaming Liang, Wenkai Chen, Xiaoyan Zhu, and Minlie Huang. 2021. A semantic-based method for unsupervised commonsense question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3037–3049. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

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.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4615–4629. Association for Computational Linguistics.

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.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3679–3686. European Language Resources Association (ELRA).

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of AI through gamification. *CoRR*, abs/2201.05320.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3878–3887. Association for Computational Linguistics.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. Semeval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 307–321. International Committee for Computational Linguistics.

Jiawei Wang and Hai Zhao. 2022. Art: All-round thinker for unsupervised commonsense question answering. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 1490–1501. International Committee on Computational Linguistics.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2022. Human parity on commonsenseqa: Augmenting self-attention with external attention. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2762–2768. ijcai.org.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1201–1207. Association for Computational Linguistics.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 93–104. Association for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4004–4010. ijcai.org.

Pei Zhou, Pegah Jandaghi, Hyundong Cho, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2021. Probing commonsense explanation in dialogue response generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4132–4146. Association for Computational Linguistics.

A Dataset Information and Statistics

The statistics of datasets in our experiments are presented in Table 6.

Dataset	N_{Inst}	N_{A}	L_Q	L_A	L_C
CSQA	1140	5	13.2	1.5	-
ARC_{E}	2376	4	19.6	3.7	-
ARC_{C}	1172	4	20.6	5.0	-
COPA	500	2	6.1	5.0	-
Swag	20005	4	12.4	11.2	-
SCT	1571	2	8.9	7.4	26.4
SQA	3525	3	11.2	5.0	19.6
CQA	6510	4	12.0	7.4	43.9

Table 6: Statistics of datasets in our experiments. N_{inst}, N_A: Number of instances and answer candidates. L_Q, L_A, L_C: Average length of the question, answer, and context.

CSQA[6] provides remarkable resources for commonsense-targeted question answering since it builds question-answer pairs based on ConceptNet. The annotators create adversarial choices based on the subgraphs in ConceptNet. Specifically, negative choices are sampled from terms related to the question in ConceptNet, making differentiating confusing for models without strong commonsense understanding.

ARC[7] is a commonsense question answering challenge that also selects phrases for science questions. The difficulty of questions is at the grade-school level and the dataset is split into the easy part (ARC_{E}) and the challenging part (ARC_{C}).

COPA[8] is a simple commonsense-targeted question answering dataset. COPA is interested in entailing a sentence by choosing a possible cause or effect of it.

Swag[9] is a large-scale commonsense question answering dataset with more than $20,000$ test data. The question is formulated as entailment that aims to satisfy the contextual integrity in commonsense.

StoryClozeTest[10] **(SCT)** is a story entailment dataset that collects 5-sentence stories with multiple ending candidates. We use the first three sentences as context and the fourth as the question.

SocialIQA[11] **(SQA)** contains questions about interactions of people in social activities. The context describes a social circumstance with related aspects, and the question asks the model to select a proper interaction.

CosmosQA[12] **(CQA)** is similar to COPA as it also asks the cause and effect of events. The difference is that CosmosQA provides an event background as the context for the question. Also, the answer of CosmosQA is longer than other datasets, which increases the difficulty for inference.

B Baselines

Answer-Level Calibration (ALC) models context-independent biases in terms of the probability of a choice without the associated context, and removes them using an unsupervised estimate of similarity with the full context. ALC consistently improves over or is competitive with baselines using standard evaluation metrics on a variety of tasks, including commonsense reasoning tasks.

Pointwise Mutual Information (PMI) factors out the probability of specific surface forms and introduce scoring-by-premise to measure the probability of the premise given the hypothesis. The authors further propose Domain Conditional PMI (PMI_{DC}) to quantify how much the premise tells us about the hypothesis within a given domain.

Self-Talk involves asking language models information-seeking questions to discover additional background knowledge. The approach improves the performance of zero-shot language model baselines on commonsense benchmarks and competes with models that obtain knowledge from external knowledge bases.

CGA is a neuro-symbolic approach to zero-shot commonsense question answering that formulates the task as inference over dynamically generated commonsense knowledge graphs. CGA generates contextually-relevant symbolic knowledge structures on demand using generative neural commonsense knowledge models, which provide interpretable reasoning paths for its predictions.

SEmantic-based Question Answering (SEQA) generates a set of plausible answers with generative models and then selects the correct choice by

[6] https://www.tau-nlp.org/commonsenseqa
[7] https://allenai.org/data/arc
[8] https://people.ict.usc.edu/ gordon/copa.html
[9] https://rowanzellers.com/swag/
[10] https://cs.rochester.edu/nlp/rocstories/
[11] https://leaderboard.allenai.org/socialiqa/submissions/public
[12] https://wilburone.github.io/cosmos/

Dataset	Question	Choices
CSQA	What island country is ferret popular?	own home, north carolina, **great britain**, hutch, outdoors
ARC_E	Which instrument measures atmospheric pressure?	**barometer**, hygrometer, thermometer, magnetometer
ARC_C	Which characteristic of a cheetah is more likely to be learned rather than inherited?	speed, a spotted coat, **hunting strategies**, claws that do not retract
COPA	The woman tolerated her friend's difficult behavior because	**the woman knew her friend was going through a hard time.** The woman felt that her friend took advantage of her kindness.
Swag	He is throwing darts at a wall.	A woman squats alongside flies side to side with his gun. A woman throws a dart at a dartboard. A woman collapses and falls to the floor. **A woman is standing next to him.**
SCT	*Rick grew up in a troubled household...* The incident caused him to turn a new leaf.	**He is happy now.** He joined a gang.
SQA	*kai was bored and had nothing to do so he played card games.* What will Kai want to do next?	do math homework, do nothing, **watch television**
CQA	*I was walking home from the store...* What may have happened to the old man?	He was waiting on the taxi. **He was waiting for the bus.** He was waiting on a ride.

Table 7: An instance from each dataset used in our experiments.

Rel.	Prompt
IsA	A is a B .
CapableOf	A is able to B .
NotCapableOf	A is unable to B .
UsedFor	A is used to B .
MadeOf	A is made of B .
PartOf	A is part of B .
HasAttribute	A is very B .
HasA	A has a B .

Table 8: Prompts used in experiments on ConceptNet.

Method	CSQA	ARC_E	ARC_C
PPL_{GPT2-M}	35.7 (0.0)	42.8 (-1.1)	27.5 (0.6)
PPL_{BERT}	42.1 (-0.3)	36.3 (-1.5)	27.1 (-0.4)
$PPL_{RoBERTa}$	45.0 (-0.7)	37.3 (-1.8)	33.2 (-0.5)
NRC	**52.3 (0.5)**	**51.9 (0.2)**	**39.8 (1.4)**

Table 9: Effect of the removal of stop words.

considering the semantic similarity between each plausible answer and each choice. SEQA achieves the best results in unsupervised settings and demonstrates stronger robustness against lexical perturbations in candidate answers. However, SEQA is a zero-shot rather than an unsupervised method because its SentenceBERT (Reimers and Gurevych, 2019) requires pre-training on NLI datasets as pointed out in (Wang and Zhao, 2022).

All-round Thinker (ArT) generates highly related knowledge by focusing on key parts in the given context in an association way, similar to human thinking, and includes a reverse thinking mechanism for causal reasoning.

C Prompts and Preprocessing

The prompts we used in experiments on ConceptNet are listed in Table 8. For $SemEval_B$, we use the prompt *"A" is not true because B.* to select an explanation for unreal commonsense expression. Prompts for question answering follow the previous configuration (Niu et al., 2021) by attaching the answer after the question.

D Normalization

Stop Word Removal For models that leverage commonsense to infer, stop words actually add noise to the inference as humans rarely use them for commonsense reasoning. Thus, we remove the scores calculated on stop words and test whether this will boost the performance of PLM-based metrics. We sample stop words from the pool provided by SpaCy to set articles and pronouns as stop words.

Shown in Table 9, NRC benefits the most from the removal of stop words, which leads to (significant) improvement on all 3 datasets. We thus conclude that NRC better takes advantage of the non-trivial components to infer.

Method	CSQA	COPA	SCT
PPL_{GPT2-M}	33.8 (-1.1)	61.0 (-7.4)	52.5 (-1.5)
PPL_{BERT}	23.0 (-7.7)	59.8 (-1.4)	59.0 (0.8)
$PPL_{RoBERTa}$	35.2 **(4.0)**	64.2 (-4.2)	65.4 (-1.9)
NRC	**43.9** (-3.5)	**74.8 (3.6)**	**81.5 (6.8)**

Table 10: Performance of conditional probability-based method. Results in bracket are the difference between **answer-based** probability.

Method	CSQA	ARC_E	ARC_C
PPL_{GPT2-M}	22.9 (-15.6)	27.8 (-16.6)	24.2 (-0.5)
PPL_{BERT}	43.2 **(2.6)**	43.6 **(6.4)**	27.8 **(1.1)**
$PPL_{RoBERTa}$	48.8 (-0.5)	40.1 (-5.4)	28.7 (-6.9)
NRC	**54.1** (0.0)	**52.1** (0.0)	**39.8** (0.0)

Table 11: Comparison among unnormalized metrics.

Conditional Method Using the conditional probability of PPL (MutualInfo-QA) is a conventional way to mitigate the lexical bias in PPL calculation (Niu et al., 2021). Namely, $\frac{p(A|Q)}{p(A)}$ is used instead of $p(A)$ for inference. $p(A)$ is divided to reduce the effect of the lexical property of the answer. We experiment with MutualInfo-QA on CSQA, COPA, and SCT datasets. For comparison, we also adapt NRC to conditional NRC by using confidence as the probability to calculate $\frac{p(A|Q)}{p(A)}$.

The results in Table 10 reflect the performance of conditional probability on three commonsense question answering datasets. Conditional NRC still outperforms other conditional metrics on all three

Method	CSQA	ARC_E	ARC_C
PPL_{GPT2-M}	32.3 (-6.2)	38.9 (-5.5)	25.8 (-3.0)
PPL_{BERT}	36.1 (8.1)	33.9 (-3.2)	24.2 (1.5)
$PPL_{RoBERTa}$	42.0 (2.2)	38.3 (-5.9)	28.0 (0.9)
NRC	50.5 (5.5)	46.0 (-1.9)	35.4 (-1.7)

Table 12: Comparison among PLMs w/ PMI. Results in bracket are the difference between **answer-based** probability.

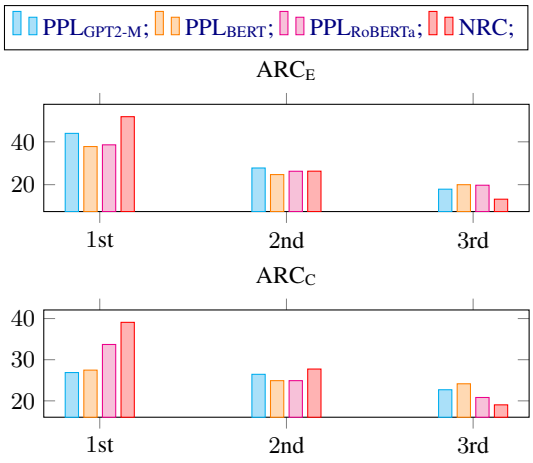

Figure 5: Ranks of PLM-based selection on easy and challenging ARC.

datasets. On COPA and SCT, NRC significantly benefits from using a conditional version, while PPL only receives a minor improvement or even a drop-down in performance. This shows the removal of initial probability is beneficial to NRC since the confidence might vary among different consistent texts. The conditional probability of NRC backfires on CSQA, which can be explained by the length (1.5 on average) of answers on CSQA datasets. As the answer is much shorter than the text used for ELECTRA pre-training, the value of $p(A)$ will add much noise to the inference. In summary, while conditional probability occasionally benefits PPL, it will benefit NRC more unless the answer text is too short.

Unnormalized Logit One way to address the constraint where candidate possibilities add up to one is to directly use the logits before applying the softmax layer. We evaluated the impact of using unnormalized logits on model performance in Table 11. The NRC model was not affected as it does not use a softmax layer. When it comes to perplexity, using unnormalized logits led to improved performance for BERT, but decreased performance for RoBERTa. For GPT2, the drop in performance was even more significant, with a decrease of over 10 points. Therefore, we can conclude that while unnormalization may enhance reasoning performance for certain types of perplexity-based reasoning, it is not a universal solution.

PMI (Holtzman et al., 2021) introduces scoring-by-premise and factors out the probability of specific surface forms to measure the probability of the premise given the hypothesis. Although PMI was initially applied only to GPT2, we wanted to investigate its impact on the reasoning abilities of different PLMs, particularly bidirectional models. As PMI regularizes the answer probability by $\frac{p(A|Q)}{p(A|Q_{domain})}$, we compare the results with answer-targeted reasoning. The results of this investigation are presented in Table 12. The analysis of the table in the article shows that PMI affects PLMs differently depending on the model architecture and the dataset used for evaluation. All PLMs show a decrease in performance in at least one dataset when PMI is applied. These results suggest that PMI's effectiveness in enhancing the reasoning abilities of PLMs may be model-dependent and dataset-dependent.

E Rank of the Choice

The accuracy only counts the matching between the golden answer and the first-rank choice. We show the ranking distribution of selected answers in Table 5 to further investigate the inference results. On the easy subsets of ARC, there does not exist a prominent advantage of NRC according to the second-rank choice rates. But when the questions become challenging, the rate of golden answers in the second rank rises, reflecting the superior capability of NRC in more challenging question answering.