# Evaluate Confidence Instead of Perplexity for Zero-shot Commonsense Reasoning

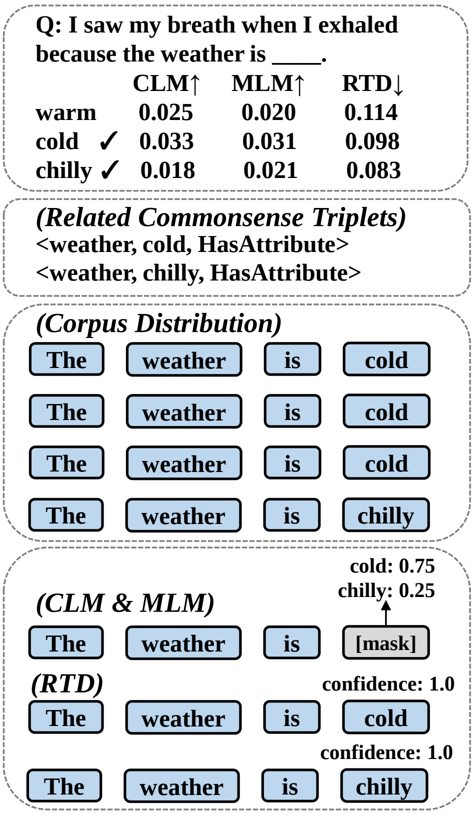
**Anonymous ACL submission**

## Abstract

Commonsense reasoning is an appealing topic in natural language processing (NLP) as it plays a fundamental role in supporting human-like actions of NLP systems. With large-scale language models as the backbone, unsupervised pre-training on numerous corpora shows the potential to capture commonsense knowledge. Current pre-trained language model (PLM)-based reasoning follows the traditional practice using perplexity metric. However, commonsense reasoning is more than existing probability evaluation, which is biased by word frequency. This paper reconsiders the nature of commonsense reasoning and proposes a novel commonsense reasoning metric, Non-Replacement Confidence (NRC). In detail, it works on PLMs according to the Replaced Token Detection (RTD) pre-training objective in ELECTRA, in which the corruption detection objective reflects the confidence in contextual integrity that is more relevant to commonsense reasoning than existing probability. Our proposed novel method boosts zero-shot performance on two commonsense reasoning benchmark datasets and further seven commonsense question-answering datasets. Our analysis shows that pre-endowed commonsense knowledge, especially for RTD-based PLMs, is essential in downstream reasoning.

## 1 Introduction

Commonsense reasoning is the underlying basis for human-like natural language understanding of machines. 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 human has been successfully applied to refine NLP systems like dialogue (Zhou et al., 2021) and generation (Chakrabarty et al., 2021).



**Q: I saw my breath when I exhaled because the weather is ____.**

| | CLM↑ | MLM↑ | RTD↓ |
|---|---|---|---|
| warm | 0.025 | 0.020 | 0.114 |
| cold ✓ | 0.033 | 0.031 | 0.098 |
| chilly ✓ | 0.018 | 0.021 | 0.083 |

Figure 1: An instance borrowed from (Niu et al., 2021) that shows the bias of PLM-based inference to high-frequency words.

As handcrafted commonsense dataset requires much time and energy from human annotators, many researchers turn to retrieving commonsense knowledge from existing language systems. Large-scale pre-trained language models (PLMs) are desirable for retrieval as they have been pre-trained on a wide variety of corpora to learn the interdependency between tokens. Petroni et al. (2019) exploit masked language modeling (MLM) strategy on BERT (Devlin et al., 2019) as a knowledge base. A series of works (Jiang et al., 2020; Alghanmi et al., 2021; Heinzerling and Inui, 2021) follow this process to prompt commonsense information from PLMs, including GPT-2 (Brown et al.,

2020) based on casual language modeling (CLM) strategy.

While MLM and CLM are the mainstream strategies for PLM-based commonsense reasoning, there still exists a doubt whether these learning objectives are competent to fully understand commonsense knowledge during pre-training. Niu et al. (2021) pointed out that the inference based on word retrieval PLMs (CLM, MLM) is likely to be biased by word frequency as presented in Figure 1. The word frequency perturbs the inference by assigning more positive scores to high-frequency words. The perturbance even leads to a wrong inference that *warm* is assigned a higher score than *chilly* in the CLM scenario.

From the view of human beings, commonsense knowledge represents facts in the physical world, whose confidence is independent of the statistical property in the corpus. The perplexity metric, biased to word frequency in the training corpus, is inconsistent with this nature. Essentially, the problem is caused by the issue that MLM and CLM constrain all sentence candidates to share a total probability of 1.0. Consequently, more frequent words will take a higher proportion of the possibility. The mutually exclusive property of perplexity underestimates confidence in other candidates when high-frequency candidates exist. On the other hand, when mentioning commonsense reasoning, we refer to confidence in the piece of knowledge rather than the existing probability of specific textual content. We thus conclude commonsense reasoning to be a discrimination rather than a generation (CLM-based generation or MLM-based prompting), which is currently done when calculating the sentence perplexity for the inference.

Based on the conclusion, we pursue a pre-trained discriminator towards better commonsense reasoning. ELECTRA (Clark et al., 2020) is a PLM trained by replaced token detection (RTD) in a GAN-like scenario. The ELECTRA discriminator is trained to detect replaced tokens from an adversarial generator. While ELECTRA does not always perform better in supervised fine-tuning (Clark et al., 2020), we find that the nature of the discriminator enables it to achieve significantly superior performance over other PLMs on zero-shot commonsense reasoning. For inference, we propose a new metric, Non-Replacement Confidence (NRC), to evaluate the integrity of fact descriptions.

We experiment with NRC on a wide variety of commonsense-related datasets. First, we evaluate the commonsense awareness of NRC on tuple and sentence-level descriptions. Then, we apply NRC to seven downstream commonsense question-answering datasets. Experiment results verify NRC to outperform perplexity-based inference by a significant gap, showing the superiority of RTD-based discriminator to capture commonsense knowledge. NRC is also efficient to calculate as it does not require mask tokens for inference.

Our analysis further discloses whether and how commonsense understanding benefits downstream inference. We gather evidence, including statistics and cases, to explain the underlying principle of the application of learned commonsense knowledge to infer. RTD-based inference is verified to be more critical to components interdependent by commonsense relationships, representing a more human-like reasoning procedure.

Our contributions are summarized as follows:

- We address the inconsistency of perplexity-based evaluation with commonsense reasoning and propose the RTD-based inference to instead evaluate the confidence.

- We implement a new RTD-based metric, NRC, which better discriminates the commonsense integrity of fact descriptions. Experiments on commonsense reasoning and question-answering verify the superiority of NRC over conventional perplexity-based inference.

- Further analyses show NRC to be more capable in not only commonsense reasoning but the application of knowledge for downstream inference as well.

## 2 Related Work

### 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 (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 physical world-based inference.

Following the storage system in databases, commonsense knowledge is generally formalized as a tuple $(LT, RT, REL)$, e.g. ConceptNet (Li et al., 2016). 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. Wang et al. create a sentence-level commonsense dataset, which validates the integrity of commonsense in real context.

### 2.2 Commonsense Reasoning with PLMs

Large-scale pre-trained language models like BERT 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 zero-shot 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 lexical properties like word frequency as perturbance to the inference. 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 Casual Language Model

GPT2 is a PLM 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. $\theta, w$ respectively refer to PLM parameters and words. The inference procedure also takes next-word prediction for perplexity (*PPL*) calculation.

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

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

### 3.2 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. The difference is MLM leverages the bidirectional context for the prediction.

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

Likewise, the inference step for MLM is revised as follows:

$$p_i = p(w_i|\text{PLM}_\theta, w_{1:i-1;i+1:n})$$

### 3.3 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 following an adversarial architecture.

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

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

We then introduce the Non-Replacement Confidence (*NRC*) metric for confidence evaluation.

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

### 3.4 Metric Comparison

*PPL* and *NRC* are both calculated based on negative log probability. While *PPL* evaluates the existing probability 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 lexical properties. 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 candidate selection), the correlation between confidence on synonym candidates is $-1$, which is contradicted to the fact that synonym supports each other for validation.

In contrast, *NRC* does not require all candidates to share the distribution but evaluates individual confidence in each candidate. Thus, there is no underlying synonym candidate that leads to an underestimation or bias toward high-frequency words. The individual evaluation also changes the correlation between synonym candidates to positive as PLMs project contextually similar components to near positions in the latent space (Devlin et al., 2019). Thus, *NRC* is a more competent metric for commonsense reasoning than *PPL*.

| Metric | Time Complexity |
|---|---|
| $\text{PPL}_{\text{CLM}}$ | $O(1)$ |
| $\text{PPL}_{\text{MLM}}$ | $O(n)$ |
| NRC | $O(1)$ |

Table 1: Time complexity of different PLM-based metrics. The complexity counts the number of PLM forwarding.

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

Table 2: 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. ($\alpha = 0.01$)

We also compare the time complexity of different metrics in Table 1. Our NRC is as efficient as the CLM-based inference since token masking is not needed to calculate the metric, which limits the efficiency of MLM-based inference.

## 4 Commonsense Reasoning

To mitigate the unfair comparison caused by the scale of parameters, this paper compares among 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 $\textbf{ELECTRA}_{\text{Large}}$[1] (24-layer, 1024-hidden size). We also include $\textbf{GPT2}_{\text{XLarge}}$ (48-layer, 1600-hidden size) for further comparison. Towards a strict unsupervised inference, we do not use any development dataset for hyperparameter selection.

### 4.1 Commonsense Probing

#### 4.1.1 Tuple-level Probing

**ConceptNet**[2] uses deep neural networks to retrieve commonsense candidates from corpus, which are validated by human annotators. Its training

[1] https://huggingface.co/google/electra-large-discriminator

[2] https://home.ttic.edu/ kgimpel/commonsense.html

dataset contains more than $600,000$ tuples with different confidences. Its test dataset requires models to discern between true commonsense tuples and adversarial fake ones.

We follow LAMA (Petroni et al., 2019) to create prompts[3] for tuples in the test dataset that can be directly represented by natural languages. Then, we differentiate the prompts by PLM-based metrics and use accuracy to evaluate the results.

Our experiment results are presented in Table 2, NRC significantly outperforms both CLM and MLM-based PPL on commonsense tuple reasoning. Considering that transformed tuple relationships are simple and unified in syntactic structures, the discriminating ability is attributed to the understanding of commonsense. Thus, the results are convincing evidence for the superiority of NRC in commonsense validation.

#### 4.1.2 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 for why the statement is against the commonsense. We continue evaluating and selecting statements and explanations according to different metrics.

As the results in Table 2, NRC is verified to perform significantly better than PPL on both differentiating and explanation, validating the superior evaluating capability of sentence-level commonsense of NRC. $\text{PPL}_{\text{RoBERTa}}$ is a competitive metric for differentiating since most statements use basic vocabulary in high frequency. Also, negative cases in SemEval are very anti-commonsense, which restrains the underestimation effect of PPL. When it comes to explanation, the gap between NRC and $\text{PPL}_{\text{RoBERTa}}$ becomes more significant since explanation requires a more complex inference ability. The comparison of sentence-level commonsense reasoning supports NRC to be a more competent metric for commonsense reasoning (differentiating and explanation) than PPL.

### 4.2 Commonsense Question Answering

For commonsense reasoning, we are interested in not only how well models understand commonsense but also how well models leverage the understanding for downstream inference. Commonsense question answering is a commonly used downstream task for the practice of commonsense understanding. We also include sampling-based baselines[5] (Self-Talk (Shwartz et al., 2020), CGA (Bosselut et al., 2021), SEQA (Niu et al., 2021)) and other strong baselines to see if NRC achieves state-of-the-art performance.

| Method | Trg | CSQA | $\text{ARC}_{\text{E}}$ | $\text{ARC}_{\text{C}}$ |
|---|---|---|---|---|
| Self-Talk | - | 32.4 | - | - |
| $\text{PPL}_{\text{GPT2-XL}}$ | A | 40.0 | 48.9 | 28.7 |
| | QA | 42.2 | 51.0 | 28.8 |
| $\text{PPL}_{\text{GPT2-M}}$ | A | 34.9 | 42.5 | 26.5 |
| | QA | 35.7 | 43.9 | 26.9 |
| $\text{PPL}_{\text{BERT}}$ | Q | 42.4 | 37.8 | 27.5 |
| | A | 30.7 | 34.8 | 25.3 |
| | QA | 35.0 | 37.2 | 24.7 |
| $\text{PPL}_{\text{RoBERTa}}$ | Q | 45.7 | 38.6 | 33.7 |
| | A | 31.2 | 33.8 | 27.7 |
| | QA | 40.0 | 37.7 | 31.9 |
| NRC | Q | 49.5 | 47.4 | 36.8 |
| | A | 47.4 | 47.3 | 37.1 |
| | QA | **51.8** | **51.7** | **38.4** |

Table 3: Experiment results on phrase selection.

#### 4.2.1 Phrase Selection

**CommonsenseQA**[6] **(CSQA)** 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 ($\text{ARC}_{\text{E}}$) and the challenging part ($\text{ARC}_{\text{C}}$).

We follow previous works (Shwartz et al., 2020; Niu et al., 2021) to calculate the metrics on different targeted components (Question (Q), Answer

[3] All prompts in our experiments can be found in Appendix B

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

[5] These methods generate many answer candidates from GPT2 to support the selection. They are more complex and time-consuming.

[6] https://www.tau-nlp.org/commonsenseqa

[7] https://allenai.org/data/arc

| Method | Trg | COPA | Swag |
|---|---|---|---|
| Self-Talk | - | 68.6 | - |
| CGA | - | 72.2 | - |
| SEQA | - | 79.4 | - |
| PPL$_{\text{GPT2-XL}}$ | A | 73.6 | 65.3 |
| | QA | 71.6 | 64.9 |
| PPL$_{\text{GPT2-M}}$ | A | 68.4 | 59.7 |
| | QA | 66.6 | 59.1 |
| PPL$_{\text{BERT}}$ | Q | 64.2 | 44.5 |
| | A | 61.2 | 63.4 |
| | QA | 64.2 | 64.1 |
| PPL$_{\text{RoBERTa}}$ | Q | 70.6 | 48.1 |
| | A | 68.4 | 71.0 |
| | QA | 75.2 | 74.5 |
| NRC | Q | **82.6** | 24.5 |
| | A | 71.2 | **77.4** |
| | QA | 78.4 | 75.4 |

Table 4: Experiment results on sentence selection.

| Method | Trg | SCT | SQA | CQA |
|---|---|---|---|---|
| Self-Talk | - | 70.4 | 47.5 | 36.1 |
| CGA | - | 71.5 | 45.4 | 42.2 |
| SEQA | - | 83.2 | 47.5 | 56.1 |
| PPL$_{\text{GPT2-XL}}$ | A | 70.6 | 41.4 | 35.5 |
| | QA | 71.5 | 41.4 | 31.1 |
| PPL$_{\text{GPT2-M}}$ | A | 54.0 | 35.6 | 27.0 |
| | QA | 55.4 | 35.4 | 18.2 |
| PPL$_{\text{BERT}}$ | Q | 63.5 | 35.7 | 32.9 |
| | A | 58.2 | 35.4 | 30.7 |
| | QA | 61.2 | 38.5 | 29.6 |
| PPL$_{\text{RoBERTa}}$ | Q | 61.5 | 37.1 | 38.6 |
| | A | 67.3 | 41.4 | 36.1 |
| | QA | 71.7 | 41.5 | 36.5 |
| NRC | Q | 65.0 | 42.8 | 41.2 |
| | A | 74.7 | 43.0 | 41.9 |
| | QA | **77.1** | **45.1** | **44.3** |

Table 5: Experiment results on context-based selection.

(A), Question+Answer (QA)) for inference. The selection results in depicted in Table 3. NRC outperforms PPL based on PLM on the same scale by a large margin (6.1, 7.8, 4.7 accuracy score), indicating NRC to be also superior in using commonsense for inference. For the easy part of ARC (ARC$_{\text{E}}$), large-scale models like GPT2$_{\text{XL}}$ seem to be able to compensate for bias in metric. However, when the questions become more challenging in ARC$_{\text{C}}$, the gap again reaches about 10.0 accuracy scores, showing the inherent differences between NRC and PPL in commonsense reasoning ability.

#### 4.2.2 Sentence Selection

**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.

Experiment results on sentence selection are presented in Table 4. NRC again shows superior performance over PPL (7.4 on COPA, 2.9 on Swag), validated by the large Swag dataset. This verifies the superiority of NRC in the application of phrase and sentence-level commonsense understanding for downstream inference. Compared to sampling-based methods, the outstanding performance of NRC also boosts state-of-the-art. The question part of Swag is not very useful for NRC probably because these questions are not dependent on the answer choices on the view of ELECTRA, which prefers to use the answer part of this dataset for inference. But when evaluating the whole question-answer pair (QA), NRC always performs better than PPL.

#### 4.2.3 Context-based Selection

**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.

As in Table 5, NRC outperforms PPL based on PLMs in the scale and the large-scale GPT2$_{\text{XLarge}}$

[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/

| Method | CSQA | $\text{ARC}_E$ | $\text{ARC}_C$ | COPA | Swag | SCT | SQA | CQA |
|---|---|---|---|---|---|---|---|---|
| $\text{PPL}_{\text{GPT2-M}}$ | 35.7 (0.0) | 42.8 (-1.1) | 27.5 (0.6) | 69.4 (1.0) | 59.3 (0.2) | 53.2 (-2.2) | 33.7 (-1.9) | 26.9 (-0.1) |
| $\text{PPL}_{\text{BERT}}$ | 42.1 (-0.3) | 36.3 (-1.5) | 27.1 (-0.4) | 66.6 (2.2) | 63.5 (-0.6) | 63.0 (-0.5) | 36.7 (-1.8) | 32.1 (-0.8) |
| $\text{PPL}_{\text{RoBERTa}}$ | 45.0 (-0.7) | 37.3 (-1.8) | 33.2 (-0.5) | 74.4 (-0.8) | 73.2 (-1.3) | 72.1 (0.4) | 41.2 (-0.3) | 38.6 (0.0) |
| NRC | 52.3 (0.5) | 51.9 (0.2) | 39.8 (1.4) | 84.2 (1.6) | 74.6 (-2.8) | 76.6 (-0.5) | 46.6 (1.5) | 44.5 (0.2) |

Table 6: Effect of the removal of stop words. Underline: The removal results in a significant improvement.

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

Table 7: 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.

| $\Delta W$ | $\text{PPL}_{\text{GPT2-M}}$ | $\text{PPL}_{\text{BERT}}$ | $\text{PPL}_{\text{RoBERTa}}$ | NRC |
|---|---|---|---|---|
| 0.00 | 35.7 | **42.4** | **45.7** | 51.8 |
| 0.25 | 35.5 | 41.8 | 45.0 | 51.9 |
| 0.50 | **35.9** | 41.2 | 44.8 | **52.2** |
| 0.75 | 35.7 | 40.6 | 44.0 | 51.7 |
| 1.00 | 35.7 | 40.2 | 43.6 | 51.7 |

Table 8: Benefits of extra weights on question concepts. **Bord:** Best performance of each PLM.

by a significant gap. On datasets with a long context (SCT and CQA), the gap becomes larger, reflecting the capability of NRC to understand the interdependency between terms in more complex contexts. On context-based selection, the sampling-based method on $\text{GPT2}_{\text{XLarge}}$ still holds state-of-the-art, which indicates that larger-scale language models still encode more knowledge in the network with much more parameters. However, the generative nature limits the understanding of the knowledge and sampling is essential to generate multiple candidates to fully retrieve the knowledge from the network. We believe that better performance and efficiency will be achieved by a larger-scale ELECTRA, which is left for future work.

## 5 Further Analysis

### 5.1 Source of Reasoning Ability

**Stop Word** 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 6, NRC benefits the most from the removal of stop words, which leads to (significant) improvement on 6 (4) out of 8 datasets. We thus conclude that NRC better takes advantage of the non-trivial components to infer.

**Synonym Replacement** We verify the advantage of NRC-based inference facing words with multiple synonyms by testing the accuracy of answer selection after synonym replacement. For implementation, we sample synonyms from Wordnet in NLTK for $10\%$ words in each question and answer text of the COPA dataset.

The results of our experiments are presented in Table 7. 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.

**Question Concept** CommonsenseQA annotates the commonsense-related phrase in each question. These phrases are connected to answer candidates in ConceptNet. For models adept at using commonsense for inference, a higher weight on the phrase should be beneficial for the inference. We thus add extra weights ($\Delta W$) and investigate the effect on different metrics.

Table 8 presents the effect of concentration on question concepts. Extra weight negatively contributes to the inference of MLM-based PLMs, indicating that they are unsuccessful in applying commonsense understanding to infer. As the negative candidates are also sampled from the neighbors of the question concept in the ConceptNet, these models are confused by ambiguity. Compared to $\text{PPL}_{\text{GPT2-M}}$, ELECTRA-based NRC benefits more from the extra weight. This verifies our claim that a discriminator better models commonsense knowledge and leverages them to infer.

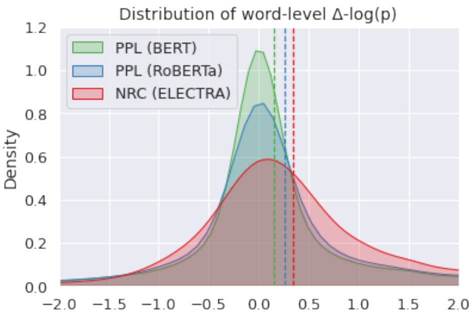


Figure 2: Distribution of the word-level differences in log probability. **Dashed line:** Average difference.

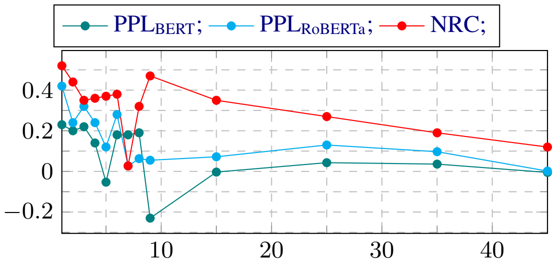


Figure 3: Relationship between word frequency and its contribution to the inference.

### 5.2 Specific Statistics

**Difference Distribution** We depict the difference distribution of log probability on COPA in Figure 2. We compare the predicted probability on the question part when it is attached by a positive or negative choice. Words are viewed as voters whose contribution to the positive choice is reflected by the difference. PPL$_{\text{GPT2-M}}$ is not included since the answer makes no difference for the question component for unidirectional PLMs. Compared to NRC, PPL difference is more likely to distribute around 0.0, indicating its lower differentiating ability. Also, the average value of NRC difference is greater than PPL difference, again supporting the stronger inference ability of NRC.

**Contribution v.s. Frequency** We continue studying the contributions of word voters. We count the frequency of words in the COPA dataset and show the relationship with their contributions in Table 3. On words with frequency $< 10$, NRC evaluation provides more positive and stable support to the right answer. The results verify our claim that NRC better evaluates the semantics of low-frequency words. The advantage of NRC over PPL decreases when the frequency rises, but NRC still holds the superiority as high-frequency words also suffer from the confidence taken by synonyms.

| Method | CSQA | COPA | SCT |
|---|---|---|---|
| PPL$_{\text{GPT2-M}}$ | 33.8 (-1.1) | 61.0 (-7.4) | 52.5 (-1.5) |
| PPL$_{\text{BERT}}$ | 23.0 (-7.7) | 59.8 (-1.4) | 59.0 (0.8) |
| PPL$_{\text{RoBERTa}}$ | 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 9: Performance of conditional probability-based method. Results in bracket are the difference between **answer-based** probability.

### 5.3 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 9 reflect the performance of conditional probability on three commonsense question-answering datasets. Conditional NRC still outperforms other conditional metrics on all three 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.

## 6 Conclusion

This paper suggests replacing perplexity with confidence to make the commonsense-targeted reasoning. We investigate the bias in the application of perplexity for inference. We propose a superior alternative, RTD-based non-replacement confidence, for better evaluation. Experiments on a wide range of commonsense reasoning and question-answering datasets provide a comprehensive analysis for the superiority of NRC.

## 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. 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 1772–1791. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5943–5959. Association for Computational Linguistics.

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.

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.

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.

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.

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.

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.

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.

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.

| Dataset | $N_{\text{Inst}}$ | $N_{\text{A}}$ | $L_Q$ | $L_A$ | $L_C$ |
|---|---|---|---|---|---|
| CSQA | 1140 | 5 | 13.2 | 1.5 | - |
| ARC$_{\text{E}}$ | 2376 | 4 | 19.6 | 3.7 | - |
| ARC$_{\text{C}}$ | 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 10: Statistics of datasets in our experiments. $N_{\text{inst}}, N_A$: Number of instances and answer candidates. $L_Q, L_A, L_C$: Average length of the question, answer, and context.

| 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 11: Prompts used in experiments on ConceptNet.

## A Dataset Statistics

The statistics of datasets in our experiments are presented in Table 10.

## B Prompts

The prompts we used in experiments on ConceptNet are listed in Table 11. For SemEval$_{\text{B}}$, 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.

## C 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 4 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.

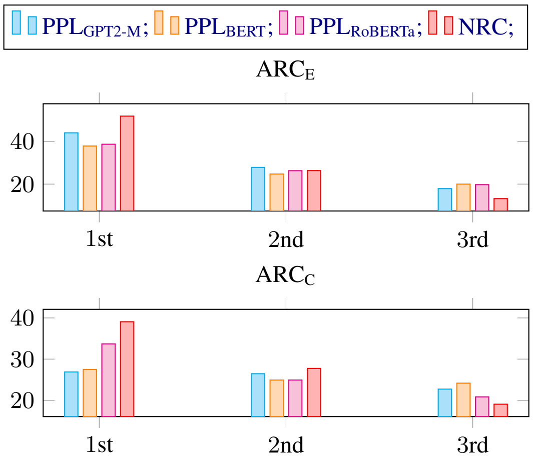


Figure 4: Ranks of PLM-based selection on easy and challenging ARC.