WINOLOGIC: A Zero-Shot Logic-based Diagnostic Dataset for Winograd Schema Challenge

Weinan He[1], **Canming Huang**[1], **Yongmei Liu**[1], **Xiaodan Zhu**[2]
[1]Dept. of Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[2]ECE & Ingenuity Labs Research Institute, Queen's University, Canada
{heweinan, huangcm}@mail2.sysu.edu.cn, ymliu@mail.sysu.edu.cn,
xiaodan.zhu@queensu.ca

Abstract

The recent success of neural language models (NLMs) on the Winograd Schema Challenge has called for further investigation of the commonsense reasoning ability of these models. Previous diagnostic datasets rely on crowd-sourcing which fails to provide coherent commonsense crucial for solving WSC problems. To better evaluate NLMs, we propose a logic-based framework that focuses on high-quality commonsense knowledge. Specifically, we identify and collect formal knowledge formulas verified by theorem provers and translate such formulas into natural language sentences. Based on these true knowledge sentences, adversarial false ones are generated. We propose a new dataset named WINOLOGIC with these sentences. Given a problem in WINOLOGIC, NLMs need to decide whether the plausible knowledge sentences could correctly solve the corresponding WSC problems in a zero-shot setting. We also ask human annotators to validate WINOLOGIC to ensure it is human-agreeable. Experiments show that NLMs still struggle to comprehend commonsense knowledge as humans do, indicating that their reasoning ability could have been overestimated.

1 Introduction

Recently, large-scale neural language models (NLMs) have shown promising results on many challenging tasks, including the Winograd Schema Challenge (WSC), a multiple-choice coreference resolution problem that is designed to test natural language understanding and reasoning with commonsense knowledge (Levesque et al., 2012; Levesque, 2014). Each problem depicts a daily situation with an ambiguous pronoun to be resolved. For example, in the WSC sentence "*the trophy doesn't fit into the brown suitcase because* ***it*** *is too large*", the pronoun "*it*" could plausibly refer to either "*the trophy*" or "*the suitcase*". The lack of training examples in WSC has driven researchers to fine-tune NLMs on similar datasets, which achieved near-human performance (Sakaguchi et al., 2020).

Still, the opacity of the NLMs has raised questions about whether they truly capture common sense or merely exploit biases. Such concerns were confirmed for the natural language inference task, as McCoy et al. (2019) discovered an LM could provide correct answers using fallible heuristics. To investigate if NLMs understand the reasons for solving WSC, Zhang et al. (2020) crowd-sourced reasons for the WSC problems and built a new dataset WinoWhy.

While in the correct direction, crowd-sourcing is far from perfect for collecting reliable explanations. Consider the following WinoWhy example.

Example 1 (WinoWhy)**.** *The trophy doesn't fit into the brown suitcase because* ***it*** *is too large. The* ***it*** *is more likely to refer to the trophy than the brown suitcase because*

(a) the brown suitcase is too large.
(b) it is a game.
(c) The trophy is not fit into the suitcase.

Given a WSC sentence with its correct answer, crowd-workers and NLMs are prompted to write a piece of text as the "reason" to justify the answer, e.g., (a), (b), and (c). Another group of crowd-workers was tasked to decide if each reason explains the answer. In Example 1, all three reasons are labeled as correct. However, upon closer inspection, *all* of them are incorrect: Reason (a) **contradicts** with the correct answer; (b) is **irrelevant**; (c) **circularly** repeats the situation. These characteristics, namely, *the lack of coherence, the lack of relevance*, and *circularity* render the explanations useless (Srinivasan and Chander, 2020).

The discouraging quality of these reasons has led us to the *logic-based* path. Compared to casually crowd-sourced justifications, we regard *reliable* commonsense knowledge as the key to solving WSC, and thus the key to constructing a diagnostic dataset. We employ first-order logic (FOL) to en-

code commonsense knowledge for two purposes: (1) FOL as a formal logic provides verifiability of the knowledge to ensure its reliability, and (2) Such knowledge serves as a better explanation since FOL enables us to see how the correct answer is derived from the problem description and the knowledge.

In this paper, we propose WINOLOGIC, a diagnostic dataset to evaluate NLMs' ability to understand and reason with commonsense knowledge for WSC. Common sense in WINOLOGIC is encoded in natural language sentences named *knowledge sentences*, based on their counterparts in FOL formulas. Consider the following example.

Example 2 (Knowledge Sentence)**.** *When someone is trying to fit an object X into a container Y, if X is too large then it wouldn't be possible to fit X into Y.*

This knowledge sentence is translated from a logical formula whose *reliability* is verified by a theorem prover. It is *coherent*, *relevant* and *unambiguous* compared to those in Example 1. Moreover, such knowledge also applies to other similar situations due to its abstract nature, while justifications such as those in Winowhy are bound to specific entities in the problem. Translating formulas into sentences not only improves the readability of the commonsense knowledge but also paves the way for evaluation. Since NLMs might not be exposed to variable symbols (X, Y, etc.) in pre-training, we provide two additional variants with variable-free knowledge sentences.

To generate reliable knowledge sentences, we use FOL to encode the knowledge and perform verifications before converting it into readable text. Specifically, we follow these steps: (1) Provide formalizations and suitable commonsense knowledge formulas for WSC problems; (2) Verify these knowledge formulas using a theorem prover; (3) Translate these formulas into *knowledge sentences* in natural language; (4) Generate adversarial false knowledge sentences with only subtle differences. WINOLOGIC is then constructed as a text classification dataset, where each problem is composed of a WSC sentence, its answer, and a plausible knowledge sentence. The task is to decide whether the knowledge supports the correct answer. Additionally, we ask human annotators to validate the dataset to ensure that the knowledge sentences are human-agreeable. Therefore the reliability of WINOLOGIC is guaranteed by both the formal verification in FOL and human validation.

WINOLOGIC, with its high-quality knowledge, is suitable for diagnosing NLMs in a zero-shot setting. Just as educators don't need large-scale examinations to test students, the small scale of WINOLOGIC doesn't invalidate its usage. What matters in tests and evaluation datasets is the quality of the problems. Focusing on the small amount of knowledge that is key for solving WSC, WINOLOGIC provides more reliable diagnostic problems. This also aligns with the few-shot or zero-shot evaluation settings, where large-scale task-specific fine-tuning is intentionally avoided. After all, people could solve WSC273 tasks without large-scale training or fine-tuning (Bender, 2015).

Experiments with three high-performing NLM architectures show that they still struggle to understand knowledge in WINOLOGIC. Even when they are fine-tuned on WinoWhy, we observe no improvement on WINOLOGIC, suggesting that they may learn little from the crowd-sourced reasons.

2 Related Work

Since its inception, WSC is gaining more and more attention. In 2016, a competition was held at the IJCAI-16 conference, but no systems qualified for the second round, as accuracies were below 60% (Davis et al., 2017). Davis (2017) collected 285 WSC problems available online, the first 273 of them are commonly referred to as WSC273. Rahman and Ng (2012) proposed the Definite Pronoun Resolution (also known as WSCR) dataset, containing 1886 problems that are considered easier than those in WSC273. Sakaguchi et al. (2020) crowd-sourced WinoGrande, a large-scale WSC-like dataset with 44k problems. An NLM-based filtering algorithm was used to identify "unbiased" problems among them. WSC problems are also incorporated into NLP benchmarks. The SuperGLUE benchmark includes a subtask SuperGLUE-WSC (804 problems), where WSC problems are cast into binary classification problems (Wang et al., 2019). While WSC273 and SuperGLUE-WSC have significant overlap, they are not entirely the same.

Recent SOTA results are achieved with fine-tuned NLMs. Two such examples are the BERT and RoBERTa models (Devlin et al., 2019; Liu et al., 2019), which have shown improvements after being fine-tuned on similar datasets. Specifically, the RoBERTa models have over 90% accuracy on SuperGLUE-WSC when fine-tuned on WinoGrande (Sakaguchi et al., 2020).

There are also analyses about the problems in WSC273 and human performance. The very first human baseline evaluation reports 92% accuracy (Bender, 2015). They noticed that for certain problems, the correct answers are only evident after the pair of questions are revealed together. Trichelair et al. (2019) discovered the *associative* subset of WSC problems where statistical correlations between the problem and the candidates could reveal the answer. Zhang et al. (2020) asked crowd-workers to classify WSC problems into different knowledge types, and they discovered spatial knowledge is particularly difficult for NLMs. Liu et al. (2020) experimented with several formulations of WSC problems and reported that task-framing has an impact on the performance of NLMs, e.g., multiple-choice setting leads to better performance.

3 Logic-based Commonsense Knowledge for WSC273

In this section, we introduce how explicit commonsense knowledge could be written in FOL and how verification is done using theorem provers. We first formalize each WSC problem into a set of logical formulas, then we provide the necessary commonsense knowledge. To ensure the correctness and validity of such knowledge, we use Z3 (de Moura and Bjørner, 2008), a theorem prover, to ensure that the formulas of knowledge could be used to derive correct answers.

Situation Calculus. We use the *situation calculus* (SC), a variant of FOL, as the representation language (Reiter, 2001). SC provides suitable constructs for modeling dynamic worlds: Entities in SC belong to either *objects*, *actions* or *situations*. Both the WSC scenario and the commonsense knowledge are represented using SC.

Example 3. *"The father couldn't lift his son because* ***he*** *was too weak."*

The scene in Example 3 before the father tried to lift his son involves the following entities:

- Objects: $Father$ and Son;
- Action: $lift(x, y)$ where x is the subject and y is the object of the action;
- Situations: S_0 is the situation where the action $lift$ has not yet happened.

To describe this scenario in SC, we use special predicates where the last argument is always a situation. For example, we could say that the son is not lifted in the initial situation S_0: $\neg Lifted(Son, S_0)$.

WSC Formalization. We formalize Example 3 with formulas:

$$\exists x. \neg Strong(x, S_0) \wedge (x \equiv Father \oplus x \equiv Son). \quad (1)$$

$$\neg Poss(lift(Father, Son), S_0). \quad (2)$$

Formula (1) says in S_0 someone was weak and that person was either the father or the son. Formula (2) states that it was not possible for the father to lift his son. To deduce that x in Formula 1 refers to $Father$, we still need suitable commonsense knowledge.

Commonsense Knowledge. Often the commonsense knowledge for WSC could be encoded using the characterizations of actions. Consider the effects and preconditions of the action of x lifts y. To express the commonsense that the subject x should be in a good physical condition for the action $lift$:

$$Poss(lift(x, y), s) \equiv Strong(x, s), \quad (3)$$

where $Poss(lift(x, y), s)$ represents the preconditions of the action.

Verification. Given the WSC scenario formalization (Formula 1 Formula 2) and the suitable commonsense knowledge (Formula 3), we use the theorem prover Z3 to formally verify the validity of commonsense knowledge. In this example, the result of Z3 indicates that the commonsense knowledge in Formula 3 supports the correct answer.

Knowledge Engineering. Each WSC problem is formalized into a set of logical formulas. With suitable commonsense knowledge, the reasoner would be able to provide correct answers. Since sophisticated automatic translation is not possible yet, we manually provide the formalization of WSC problems and the commonsense knowledge through knowledge engineering, relying on experts that are fluent in FOL. We collect all the commonsense knowledge formulas that are verified by Z3.

4 Knowledge Sentences and WINOLOGIC

In this section, we present the creation of positive and negative knowledge sentences as in Figure 1, then we describe the resulting WINOLOGIC.

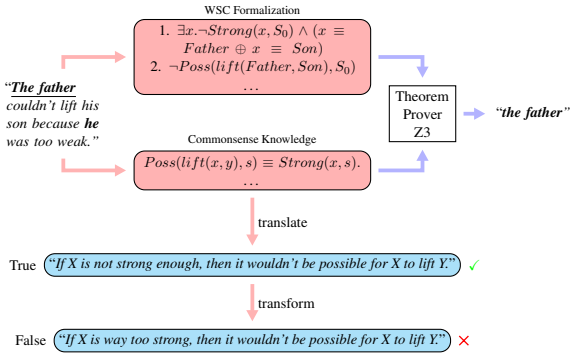

Figure 1: The process of generating both true and false knowledge sentences. We first create the formalization of WSC problems and provide the commonsense knowledge formulas. Together they are verified by a theorem prover to be able to reason the correct answer. Then we translate the commonsense into natural language sentences, where the false ones are derived.

4.1 Knowledge Sentences with Variables

To better evaluate NLM's ability to understand commonsense, we transform the logical knowledge formulas into natural language sentences.

For each WSC problem, we pick the essential commonsense knowledge formulas and translate them into natural language sentences. For example, the formula $Poss(lift(x, y), s) \equiv Strong(x, s)$ translates to the knowledge sentence "*When person X is about to lift person Y up, if X is not strong enough, then it wouldn't be possible for X to lift Y*". We adhere to rules in translation to preserve coherence:

1. The variables (of sort objects) x, y, z in the formulas are preserved. By doing so, sentences retain a certain level of abstraction. This not only aligns with the idea that knowledge should be widely applicable but also provides the necessary means to reduce the ambiguity in sentences involving multiple people, objects, etc.
2. The symbols of the predicates are usually retained in the sentence whenever it is possible and natural to do so.
3. In some cases, extra information is added to the knowledge sentences, such as the type information "person", "object", etc.

For each true knowledge sentence, we generate a false one that is still plausible. These false knowledge sentences should be 1) *relevant* to the WSC problem as much as possible and 2) obviously wrong; That is, false knowledge sentences will not help the reasoning process. For relevance, we adopt one of the following transformations for each knowledge sentence.

Negation. *Negating* the meaning of the sentence will create a contradicting statement, following the Natural Logic inference system of Angeli and Manning (2014). For example, "*X is tall enough*" is transformed into "*X is not tall enough*".

Swapped or Replaced. *Swap* the positions of the two parties either in the antecedent or the consequent. Or *replace* one with the other. For example, "*X is larger than Y*" becomes "*Y is larger than X*"; and "*X is angry*" becomes "*Y is angry*".

Changed. *Change* the content of the sentence while preserving relevance. For example, "*person X is a suspect of criminal*" becomes "*person X is hurt by a criminal*".

Others. Use multiple transformations.

We manually provide a pair of true and false knowledge sentences based on the verified logical formulas for each WSC problem, but for some problems, more than one knowledge formula is considered. In the end, we obtain a total of 562 knowledge sentences, half of which are true, and the other half false. We denote this set of knowledge sentences as the *variable* set.

4.2 Grounded and Natural Knowledge Sentences

To better understand how well NLMs handle commonsense, we also provide two more sets of knowledge sentences, the *grounded* and the *natural* sets, based on the variable set.

Grounded. In the grounded set, the occurrences of variables are substituted with their corresponding mentions, the noun phrases in the WSC problem. E.g., "*it wouldn't be possible for the father to lift the son*".

Natural. To generate sentences that are more natural while preserving the unambiguous nature, variables in the sentences are removed. If there are multiple parties, we use ordinal numbers to differentiate between them. For example, if a sentence involves multiple persons, we use "the first person", "the second person" etc. E.g., "*When someone is trying to fit an object into a container, if the object is too large then it wouldn't be possible to fit the object into the container.*"

Variant	Jaccard Similarity	Edit Distance	Length Difference
Variable	0.9325	2.4626	0.3203
Grounded	0.9270	2.6050	0.5125
Natural	0.9178	2.5730	0.4769

Table 1: Statistics of knowledge sentences; Average Jaccard similarity coefficient, average edit distance, average length difference are reported between a pair of true and false knowledge sentences. All values are reported on the token level.

4.3 Analysis of Knowledge Sentences

Table 1 shows the statistics of the knowledge sentences. The Jaccard similarity coefficient measures the overlap of words between the two sentences. These numbers imply that the similarity between the true and false knowledge sentences is rather high, which comes as no surprise as the false sentences are generated by design to be close to the true ones.

4.4 WINOLOGIC and Validation

We concatenate the WSC sentence, its correct answer, and a plausible knowledge sentence, resulting in a binary classification problem in WINOLOGIC. Example 4 shows a pair of WINOLOGIC-Natural problems corresponding to the same WSC problem, differing only in the small change of a few words. This also reflects the designing characteristics of WSC.

Example 4. *The man couldn't lift his son because he was so weak. The he is more likely to refer to the man than the son because*

(a) when a person is about to lift another person up, if the first person is not strong enough, then it wouldn't be possible for the first person to lift the second person. ✓

(b) when a person is about to lift another person up, if the first person is way too strong, then it wouldn't be possible for the first person to lift the second person. ✗

We construct three variants of WINOLOGIC from the three sets of knowledge sentences: (1) WINOLOGIC-variable, (2) WINOLOGIC-grounded, and (3) WINOLOGIC-natural. Each variant contains 562 binary classification problems.

In addition to using formally verified knowledge formulas, we also conduct human validations on the original WINOLOGIC-variable variant. Six undergraduate students[1] are asked to decide if the knowledge sentence adheres to commonsense and if it is valid to support the answer. Between annotators and the ground labels, the average raw percentage agreement is 93.86% while the average Cohen's kappa coefficient is 0.8772, showing that the WINOLOGIC knowledge is rather human-agreeable. Among the 562 problems, 518 (92.17%) are deemed valid as at least 5 out of 6 annotators agree with the ground labels. It contains 249 problems that are labeled true and 269 false. In the next section, we use this human-validated subset of WINOLOGIC for evaluation.

5 Evaluation

In this section, we describe baseline implementations for WINOLOGIC and present the evaluation results. The dataset, code, and hyper-parameters are available in the supplementary materials.

5.1 Neural Language Models

We consider the three transformer-based NLM architectures: GPT-2, BERT, and RoBERTa, as they show promising results on WSC (Kocijan et al., 2020). In addition to their unique training objectives, we also utilize the sequence-to-sequence output of these architectures.

1. **GPT-2** (Radford et al., 2019): GPT-2 was shown to achieve 70.7% of accuracy on WSC273 (Radford et al., 2019). We employ the pre-trained objective of GPT-2 that was designed for traditional language modeling tasks.
2. **BERT** and **RoBERTa** (Devlin et al., 2019; Liu et al., 2019): We utilize the *masked language modeling* and *sequence classification* pre-trained objectives. BERT was shown to achieve comparable accuracy as GPT-2, while RoBERTa has maintained the SOTA performance on WSC.

All vanilla pre-trained models are obtained from the `Transformers` library (Wolf et al., 2020). Besides the sequence-to-sequence NLMs, we also utilize the LM heads provided by the library. These LM heads take the input from the NLM and return desired results according to the pre-training task objective. For example, the `BertForMaskedLM` returns the prediction scores for the masked tokens along with BERT sequence-to-sequence output.

As fine-tuning shows promising results on WSC, we also fine-tuned BERT and RoBERTa on WSCR

[1]For more detail about the annotators, please refer to the Appendix.

Table 2: Accuracies of 14 NLMs using the first two baseline methods on 3 variants of WINOLOGIC. Values in parentheses for baseline 2 are standard deviations. "Majority" is a voting ensemble of the NLMs. WR=WSCR, WG=WinoGrande.

NLM	Baseline 1			Baseline 2		
	Variable	Grounded	Natural	Variable	Grounded	Natural
Majority	54.63%	**57.53%**	53.67%	50% (0.0087)	49.34% (0.0063)	49.03% (0.0042)
BERT (base)	50.97%	54.44%	51.54%	50.46% (0.0111)	48.69% (0.0075)	49.42% (0.0076)
BERT (large)	50.77%	54.25%	49.61%	49.42% (0.0089)	48.88% (0.01)	49.54% (0.0169)
RoBERTa (base)	50.58%	54.05%	51.54%	49.58% (0.0077)	49.77% (0.0023)	49.38% (0.0116)
RoBERTa (large)	52.12%	55.98%	52.12%	47.84% (0.0043)	47.61% (0.0072)	48.53% (0.0076)
GPT-2 (base)	51.35%	47.88%	52.32%	49.31% (0.0083)	48.92% (0.0065)	48.92% (0.0079)
GPT-2 (large)	50.39%	51.74%	51.74%	50.39% (0.0156)	49.03% (0.0116)	49.46% (0.0143)
BERT (base) + WR	49.23%	51.93%	50.58%	48.88% (0.0087)	47.88% (0.0129)	48.26% (0.0118)
BERT (large) + WR	51.16%	52.7%	49.61%	48.65% (0.0127)	49.31% (0.0097)	48.92% (0.0113)
RoBERTa (base) + WR	52.12%	53.28%	49.61%	49.38% (0.0077)	49.07% (0.0114)	49.73% (0.013)
RoBERTa (large) + WR	52.9%	56.76%	55.21%	48.26% (0.0047)	48.11% (0.0072)	48.46% (0.004)
BERT (base) + WG	51.35%	55.98%	51.35%	49.92% (0.0066)	48.84% (0.0061)	50% (0.0142)
BERT (large) + WG	50.39%	54.25%	51.74%	50.42% (0.0114)	49.81% (0.0106)	49.23% (0.0093)
RoBERTa (base) + WG	51.35%	54.05%	51.35%	49.58% (0.0109)	49.31% (0.0111)	49.46% (0.0097)
RoBERTa (large) + WG	**55.79%**	57.14%	**55.98%**	52.66% (0.0318)	53.98% (0.0389)	52.16% (0.0202)

or WinoGrande (Kocijan et al., 2019; Sakaguchi et al., 2020). Additionally, we use two natural language inference (NLI) datasets, MNLI and QNLI, for fine-tuning (Wang et al., 2018; Williams et al., 2018). "BERT (large) + WR" denotes the large BERT model fine-tuned on WSCR. We first introduce two baseline methods inspired by Zhang et al. (2020).

5.2 Baseline 1: Using Pre-trained Objectives

GPT-2. Given a WINOLOGIC problem, we first tokenize it into a sequence of token $t_1, t_2, \ldots, t_n$ as the input. The language modeling head (`GPT2LMHead`) predicts the tokens from start to end. We only calculate the cross-entropy values starting from the token m, where m is the first token immediately after the unknown pronoun in the WSC sentence. This partial prediction scheme turns out to have better performance than predicting the whole sentence (Zhang et al., 2020). In Example 4, the tokens starting from "*was*" are predicted and the corresponding cross-entropies are calculated.

BERT and RoBERTa. For BERT and RoBERTa, we use the masked token prediction objective. In the token sequence $t_1, t_2, \ldots, t_n$ for a given WINOLOGIC problem, we mask the candidate tokens and use the masked LM head to predict them. For example, a WINOLOGIC problem contains the answer part of form: "... *the* $\langle pronoun \rangle$ *is more likely to refer to* $\langle cand1 \rangle$ *than* $\langle cand2 \rangle$ *because* ..." where

- $\langle pronoun \rangle$ is the unknown pronoun,
- $\langle cand1 \rangle$ is the correct reference and
- $\langle cand2 \rangle$ is the incorrect reference.

Both $\langle cand1 \rangle$ and $\langle cand2 \rangle$ are masked and the masked LM heads are used to predict the cross-entropy losses for them.

NLM	Baseline 3		
	Variable	Grounded	Natural
Majority	52.32%	53.67%	51.35%
BERT (base) + QNLI	50.97%	51.54%	49.23%
RoBERTa (base) + QNLI	53.67%	48.65%	52.7%
BERT (base) + MNLI	51.93%	51.74%	50.58%
RoBERTa (base) + MNLI	51.74%	52.32%	51.93%
RoBERTa (large) + MNLI	51.93%	54.63%	52.32%

Table 3: Accuracies of 5 NLMs using the third baseline method on 3 variants of WINOLOGIC. "Majority" is a voting ensemble of the NLMs.

Classifying WINOLOGIC Problems For a WINOLOGIC problem, if the knowledge sentence is correct, it should support the correct answer "*the* $\langle pronoun \rangle$ *is more likely to refer to* $\langle cand1 \rangle$ *than* $\langle cand2 \rangle$", instead of the incorrect one where $\langle cand1 \rangle$ and $\langle cand2 \rangle$ are swapped. Thus, we create two texts $s_{original}$ and s_{swap} where the former is the same as the original WINOLOGIC problem while the latter swaps the two candidates. By comparing the cross-entropy values of the two texts, an NLM determines whether the knowledge sentence is correct. Compared to the baseline used for WinoWhy, this method doesn't need to find a threshold of probabilities and thus is unsupervised.

Model	Variable-True	Variable-False
Baseline1	52.24% (0.0240)	50.74% (0.0102)
Baseline2	72.31% (0.1122)	28.63% (0.1079)
Baseline3	15.50% (0.1784)	85.87% (0.1508)
Model	**Grounded-True**	**Grounded-False**
Baseline1	56.22% (0.0321)	51.73% (0.0219)
Baseline2	85.37% (0.0994)	15.77% (0.0859)
Naseline3	28.51% (0.2621)	73.31% (0.2724)
Model	**Natural-True**	**Natural-False**
Baseline1	52.58% (0.0287)	50.96% (0.0162)
Baseline2	67.42% (0.1993)	32.70% (0.1820)
Baseline3	35.42% (0.3544)	66.10% (0.3300)

Table 4: Average accuracies of NLMs in each baseline methods on True and False WINOLOGIC examples

5.3 Baseline 2: Using Auxiliary Classifier

To investigate whether the reasons in WinoWhy could potentially help NLMs to capture knowledge needed in WINOLOGIC, we retain the baseline method used for WinoWhy. Given a WINOLOGIC problem, the NLM outputs a sequence of hidden states where the linear classifier learns to determine if the sentence is true or false. We take the 2865 WinoWhy examples to fine-tune the composite model.

5.4 Baseline 3: Using Sequence Classification

As shown in Wang et al. (2018), NLMs have impressive performances on NLI tasks, which inspire us to cast WINOLOGIC problems as NLI problems. Specifically, we utilize the sequence classification heads of BERT and RoBERTa to decide whether the knowledge sentence could entail the WSC sentence with the unknown pronoun replaced by the correct answer. We evaluate the models fine-tuned on either QNLI or MNLI.

5.5 Result Analysis

We observe from Table 2 and Table 3 that NLMs are struggling on WINOLOGIC, as none have accuracies above 58% on all three variants, while the best performance is achieved by the larger RoBERTa model fine-tuned on WinoGrande. Comparing the performances between the three baselines, the first one outperforms the other two. This implies the justifications in WinoWhy are not useful for the NLMs. On the other hand, fine-tuning NLMs on larger WSC-like datasets does have a positive result, especially for the larger models. We also observe that the performances on the three variants of WINOLOGIC are rather similar, suggesting that although they use different mentioning schemes (variables, grounded, and ordinal mentions), the difficulty doesn't come from the use of the more abstract variables.

From Table 4, we take a closer look at how NLMs handle the true and false subsets of WINOLOGIC (249 and 269 problems respectively). While the first baseline has similar accuracies on the two subsets, the other two have clear tendencies. NLMs in the second baseline are fine-tuned on WinoWhy, and they favor classifying WINOLOGIC problems as positive. One of the possible explanations could be that WINOLOGIC examples, no matter true or false, resemble those positive ones in WinoWhy. On the other hand, the methods in baseline 3 are clearly rejecting WINOLOGIC examples. The stark difference in task setting may contribute to this phenomenon.

In Table 5, we present the average performances of NLMs in each baseline on the different types of false WINOLOGIC-Variable. Accuracies between different types of false knowledge sentences are similar.

6 Discussion

Although crowd-sourcing is a valuable tool to put collective intelligence to work, it may not be an ideal method to produce reliable explanations for WSC. Firstly, generating explanatory text for even simple problems is already immensely more difficult than classifying its correctness. People often apply common sense subconsciously, and thus could potentially trivialize its significance. For example, we often take the closed-world assumption for granted. Secondly, it requires a deep reasoning process for solving many WSC problems. In this case, precision and unambiguity would be vital, which is not exactly the strong suit of crowd-sourcing. Last but not least, from our prior experience in collecting expert-provided justifications similar to those in WinoWhy, we notice the same issues (e.g., circularity and incoherence) in the collection, even though these NLP-experts were asked to meticulously explain the reason for WSC problems. In this paper, we leverage the unambiguity of FOL which enables us to precisely represent the commonsense knowledge, while theorem provers are used to verify its validity, addressing the major issues in crowd-sourcing explanations.

Model	Negation (40)	Swapped (72)	Replaced (117)	Changed (20)	Other (20)	False (269)
Baseline1	53.75% (0.043)	51.59% (0.031)	49.33% (0.0293)	48.93% (0.0849)	51.79% (0.0586)	50.74% (0.0102)
Baseline2	30.57% (0.1068)	24.54% (0.1123)	28.38% (0.1115)	38.36% (0.1446)	31.21% (0.0848)	28.63% (0.1079)
Baseline3	87.5% (0.1245)	86.39% (0.1304)	84.79% (0.1758)	90% (0.1095)	83% (0.2015)	85.87% (0.1508)
Human	97.92%	96.06%	97.72%	97.5%	93.33%	96.96%

Table 5: Average accuracies of 3 baseline methods on False examples in WINOLOGIC-Variable

Although expert-sourcing formal knowledge is expensive, quality is of utmost importance, especially in a zero-shot evaluation setting. WINOLOGIC identifies the essential commonsense knowledge in WSC273, and thus serves as a better diagnostic dataset for testing current NLMs. The zero-shot evaluation setting challenges systems to perform commonsense reasoning without resorting to fine-tuning, towards reaching human-level capabilities.

Moreover, it is feasible to extend our approach to other reasoning tasks. Similar to real-life tests where teachers only design a small number of high-quality problems, it is not necessary for a diagnostic dataset to contain tens of thousands of problem instances. After identifying the core of the task and choosing a proper representation, the cost of expert-sourcing is manageable by limiting the number of instances to be within an acceptable range. As long as typical examples are included and the core to the reasoning tasks is targeted, a diagnostic dataset suffices without large-scale problem instances.

While NLMs benefit from large-scale pre-training and fine-tuning, their performances on WINOLOGIC fall short of expectations. The "common sense" they capture is not yet robust enough for them to understand the knowledge behind WSC. The contrasting disparity of performances further confirms the possibility of "overestimating the true capabilities of machine intelligence on common sense reasoning" (Sakaguchi et al., 2020), as systems should not overlook the subtle difference between a pair of true and false knowledge sentences. In Example 4, human annotators have no trouble differentiating between "not strong enough" and "way too strong", while NLMs are confused. It is therefore vital for WSC solvers to adopt knowledge-aware mechanisms that are capable of integrating common sense beyond shallow inference.

7 Conclusion

We show a knowledge-based perspective for evaluating recent improvements on the Winograd Schema Challenge. Instead of crowd-sourcing justifications, we take a more reliable logic-based route. By formalizing the WSC problems and the needed commonsense knowledge into FOL formulas, we verify the capability of using such knowledge for solving the problems. To test recent neural language models, we translate the commonsense knowledge into natural language sentences. Simple transformations create negative versions of these sentences. We propose WINOLOGIC, a novel evaluation dataset of 562 WSC-related knowledge classification problems that are also human-validated. This new task requires systems to determine whether a knowledge sentence could be used to solve the corresponding WSC problem. The experiments show that recent language models do not have a correct understanding of the required knowledge, even though they are already fine-tuned on the similar WinoWhy dataset. Our exploration suggests that the challenge of commonsense reasoning in WSC is still bottlenecked by the lack of machine common sense.

Acknowledgement

We appreciate the fruitful discussion with Dr. Zhanhao Xiao. We acknowledge support from the Natural Science Foundation of China under Grant No. 62076261.

References

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 534–545, Doha, Qatar. Association for Computational Linguistics.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *Proceedings of the 26th Modern AI and Cognitive Science Conference 2015, Greensboro, NC, USA, April 25-26, 2015.*, pages 39–45.

Ernest Davis. 2017. Logical formalizations of commonsense reasoning: A survey. *J. Artif. Int. Res.*, 59(1):651–723.

Ernest Davis, Leora Morgenstern, and Charles L. Ortiz Jr. 2017. The first winograd schema challenge at IJCAI-16. *AI Magazine*, 38(3):97–98.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4837–4842.

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

Hector J. Levesque. 2014. On our best behaviour. *Artif. Intell.*, 212(1):27–35.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*.

Haokun Liu, William Huang, Dhara Mungra, and Samuel R. Bowman. 2020. Precise task formalization matters in Winograd schema evaluations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8275–8280, Online. Association for Computational Linguistics.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Leonardo Mendonça de Moura and Nikolaj Bjørner. 2008. Z3: an efficient SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, TACAS 2008, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2008, Budapest, Hungary, March 29-April 6, 2008. Proceedings*, volume 4963 of *Lecture Notes in Computer Science*, pages 337–340. Springer.

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

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Raymond Reiter. 2001. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT press.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Ramya Srinivasan and Ajay Chander. 2020. Explanation perspectives from the cognitive sciences—a survey. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4812–4818. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3373–3378, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5736–5745. Association for Computational Linguistics.

Appendix

Computing Infrastructure

We use a workstation with

- Intel i9-10980XE (16 cores) CPU;
- 64GB of RAM;
- 2 GPUs (RTX 3090).

Hyper-parameters

For the second baseline method, we manually set the hyper-parameters shown in Table 6.

Learning Rate / Seed / Model	42	1718	3149	8747	9334
BERT(base)	1e-2	1e-2	5e-3	1e-2	1e-2
BERT(large)	1e-4	1e-4	1e-4	1e-4	1e-4
RoBERTa(base)	2e-2	3e-2	2e-2	1e-2	2e-2
RoBERTa(large)	2e-5	2e-5	1e-5	2e-5	2e-5
GPT-2(base)	5e-4	5e-4	5e-4	2e-4	5e-4
GPT-2(large)	1e-4	1e-4	1e-4	1e-4	1e-4
BERT(base)+WSCR	1e-2	1e-2	1e-2	1e-2	1e-2
BERT(large)+WSCR	1e-4	5e-5	1e-4	1e-4	1e-4
BERT(base)+WinoGrande	1e-2	1e-2	1e-2	1e-2	5e-3
BERT(large)+WinoGrande	5e-5	1e-4	1e-4	5e-5	1e-4
RoBERTa(base)+WSCR	1e-2	2e-2	1e-2	5e-3	2e-2
RoBERTa(large)+WSCR	2e-5	2e-5	2e-5	2e-5	2e-5
RoBERTa(base)+WinoGrande	2e-2	1e-2	5e-3	1e-2	1e-2
RoBERTa(large)+WinoGrande	2e-5	1e-5	2e-5	2e-5	1e-5

Table 6: Hyper-parameters for Experiment "Using Auxiliary Linear Classifier". We randomly select five seeds: 42, 1718, 3149, 8747, and 9334. Learning rates corresponded to each seed for each model are shown in the table. The number of epochs for all models is set to 30. Batch size for base models except GPT-2(base) is 32, while the batch size for large models and GPT-2(base) is 1.

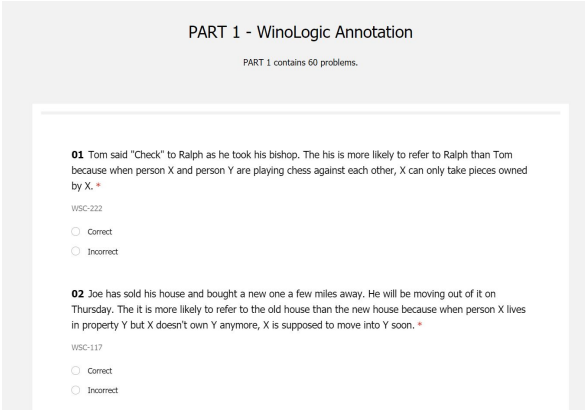

Figure 2: Annotation Interface: An annotator decides whether a WINOLOGIC problem is correct or not.

Description of Task and Instructions

- A WSC sentence describes a daily situation, containing an unknown pronoun that could refer to one of two plausible mentions. The task it to resolve the unknown pronoun is called coreference resolution.
  - [Example] **The city councilmen** refused **the demonstrators** a permit because *they* feared violence.
- WinoLogic problems are text classification problems related to WSC problems. It contains over 500 problems. Each problem consists of two sentences.
  - The first sentence is the WSC sentence
  - The second sentence 1) resolves correctly the unknown pronoun in the WSC sentence and 2) tries to provide commonsense knowledge needed for the coreference resolution problem.
    - [Example 2] The they is more likely to refer to the city councilmen than the demonstrators because *when people X applies for a permit to demonstrate, if administrators Y denies it, then X is more likely to fear violence.*
    - The sentence in Example 2 points out *they* refers to *the city councilmen*, and provides a plausible commonsense knowledge sentence, which could be correct or incorrect.
  - For each problem in WinoLogic, classify it into either **correct** or **incorrect**.
    - **Correct**: The knowledge sentence could be used to solve the WSC coreference problem, and it it correct w.r.t. commonsense.
    - **Incorrect**: The knowledge sentence **could not** be used to solve the WSC coreference problem, or it it **incorrect** w.r.t. commonsense.

Your task is to read each WinoLogic problems carefully, and classify them accordingly.

Figure 3: Annotation Instruction

Annotations

Human annotations are done using an online questionnaire system where each WINOLOGIC problem is presented as a binary classification problem. Figure 2 shows the screenshot of the annotation interface. Before the annotations, the annotators received instructions in a PDF file with a screenshot shown in Figure 3.

Annotator Details Each annotator was recruited with their consent to assist research. They receive fair compensation of 500 Chinese Yuan per annotator. The annotation was not mandatory coursework. None of the annotators received course credit for it. From the submitted data, they worked for an average of 8.7 hours. Note that we didn't pose restrictions on how long they should spend on annotating.