Diagnosing the First-Order Logical Reasoning Ability Through LogicNLI

Jidong Tian[1,2†], **Yitian Li**[1,2†], **Wenqing Chen**[1,2], **Liqiang Xiao**[1,2], **Hao He**[1,2‡] and **Yaohui Jin**[1,2]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

[2]State Key Lab of Advanced Optical Communication System and Network, Shanghai Jiao Tong University

{frank92, yitian_li, wenqingchen, xiaoliqiang, hehao, jinyh}@sjtu.edu.cn

Abstract

Recently, language models (LMs) have achieved significant performance on many NLU tasks, which has spurred widespread interest for their possible applications in the scientific and social area. However, LMs have faced much criticism of whether they are truly capable of reasoning in NLU. In this work, we propose a diagnostic method for first-order logic (FOL) reasoning with a new proposed benchmark, LogicNLI. LogicNLI is an NLI-style dataset that effectively disentangles the target FOL reasoning from commonsense inference and can be used to diagnose LMs from four perspectives: accuracy, robustness, generalization, and traceability. Experiments on BERT, RoBERTa, and XLNet, have uncovered the weaknesses of these LMs on FOL reasoning, which motivates future exploration to enhance the reasoning ability.

1 Introduction

Recently, Transformers-based (Vaswani et al., 2017) language models (LMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have achieved great success on natural language understanding (NLU). However, there are growing concerns about whether LMs can truly understand natural language or not. Tasks with complex reasoning have provided evidence that LMs lack expected reasoning abilities (Liu et al., 2020; Bhagavatula et al., 2020). Even if neural models can make correct predictions, they tend to make decisions through spurious statistical correlations rather than reasoning abilities (Kaushik and Lipton, 2018; Ribeiro et al., 2019; Jiang and Bansal, 2019; McCoy et al., 2019). Therefore, an increasing number of studies have focused on diagnosing specific reasoning abilities of state-of-the-art LMs (Sugawara et al., 2020; Gontier et al., 2020).

First-order logical (FOL) reasoning is one of the most widely used reasoning forms in natural language (Davis, 2017; Yu et al., 2020), which has a simple paradigm consisting of combinations of seven fundamental logics (FOLs, including conjunction $\wedge$, disjunction $\vee$, negation $\neg$, implication $\rightarrow$, equation $\equiv$, universal quantifier $\forall$, and existential quantifier $\exists$) with simple propositions (Davis, 2017). Nevertheless, whether LMs can truly make FOL reasoning is still inconclusive in NLP (Hahn et al., 2021; Clark et al., 2020).

As a result, we propose a systematic diagnostic method for FOL reasoning by proposing a novel benchmark, named **Logic**al **N**atural **L**anguage **I**nference (LogicNLI). The proposed benchmark follows three principles: 1) It includes abundant logical expressions covering all seven FOLs and their commonly used combinations in texts; 2) The instances of the benchmark conform to natural language; 3) It introduces as little commonsense as possible to prevent the targeting FOL reasoning and commonsense inference from being entangled with each other (Clark et al., 2020). According to the principles, LogicNLI is an NLI-style dataset (Bowman et al., 2015; Talmor et al., 2020), including triplets of facts, rules, and a statement. The objective is to determine the logical relation (entailment, contradiction, or neutral in NLI (Bowman et al., 2015)) between the premise (facts and rules) and its corresponding hypothesis (statement) by FOL reasoning shown in Figure 1. In practice, we have introduced an additional logical relation, "Paradox", to represent the situation where the hypothesis and its negative proposition can be logically entailed to the premise simultaneously based on different reasoning paths (bottom of Figure 1). This novel logical relation forces the model to search at least two reasoning paths to infer the authenticity of two opposing propositions, thereby effectively avoiding spurious correlations caused by dataset bias.

Based on LogicNLI, we propose a systematic

[†] These authors contributed equally.

[‡] Corresponding author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3738–3747

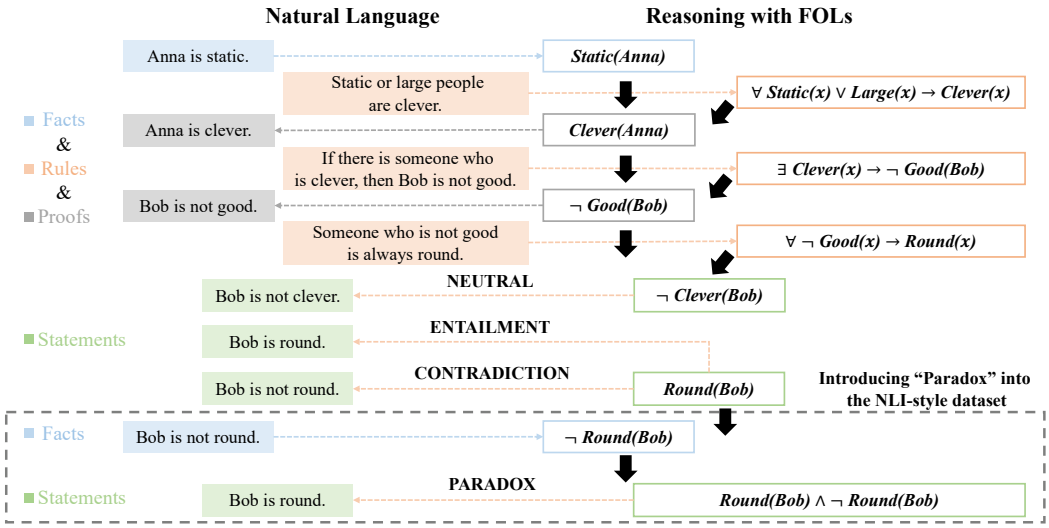

Figure 1: Reasoning processes in LogicNLI. Given a set of facts (Blue) and rules (Orange), The first step is to translate the language expressions into the FOL expressions. Based on expressions, logical reasoning is made step by step, where proofs (Grey) are the intermediate results of each step. Finally, the proposed statements (Green) are judged based on multi-step reasoning. Besides, LogicNLI provides a new condition of "PARADOX" that both the positive and negative propositions can be inferred simultaneously (shown in the dotted frame).

diagnostic approach by comprehensively considering four perspectives: accuracy, robustness to irrelevant information, more-hop generalization, and proof-based traceability. We perform diagnosis on three state-of-the-art LMs, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). Results reveal that LMs can neither fully understand the logical rules nor apply them to reason like humans. In conclusion, our main contributions include: 1) We design a novel benchmark, LogicNLI, following three basic principles to diagnose LMs' FOL reasoning ability. This method of benchmark construction is general for different reasoning types in NLU. 2) Based on LogicNLI, we design a diagnostic approach composed of accuracy, robustness, generalization, and traceability, which measures LMs' FOL reasoning ability from different perspectives. 3) Results on three LMs show that even the best performing model on LogicNLI, RoBERTa, cannot fully infer according to logic and generalize to different scenarios. Analysis could inspire the further exploration of incomprehensible logic.

2 Related Work

2.1 NLU Benchmark

With the development of language models, many traditional NLU datasets, such as SQuAD (Rajpurkar et al., 2016, 2018), HotpotQA (Yang et al., 2018), and MNLI (Williams et al., 2018), seem to have been resolved. However, new concerns about spurious correlations (Ribeiro et al., 2019; Jiang and Bansal, 2019) motivate novel datasets to benchmark specific NLU abilities. Some of these datasets concentrated on commonsense or domain knowledge, such as CosmosQA (Huang et al., 2019), PiQA (Bisk et al., 2020), CommonsenseQA (Talmor et al., 2019), and SocialIQA (Sap et al., 2019). Other datasets focused on specific reasoning in NLU, including numerical reasoning (Amini et al., 2019; Dua et al., 2019; Tafjord et al., 2019; Ravichander et al., 2019), coreferential reasoning (Dasigi et al., 2019; Sakaguchi et al., 2020), abductive reasoning based on commonsense (Bhagavatula et al., 2020), and pragmatic reasoning that is originated from linguistics (Jeretic et al., 2020). These studies provided diverse views to benchmark how machines understand language.

2.2 FOL Reasoning Benchmark

Among these NLU abilities, FOL reasoning is a fundamental reasoning ability that attracts an increasing number of studies to benchmark. LogiQA (Liu et al., 2020) and ReClor (Yu et al., 2020) are two comprehensive datasets with domain knowledge. However, even if a model performs poorly on these datasets, it is inconclusive that the model lacks the FOL reasoning ability because the targeting ability cannot be disentangled from other reasoning abilities, such as commonsense in-

Dataset	Logic		Natural	Commonsense	
	#FOLs	Proof	Language	Domain	Predicate
LogiQA	5	×	√	√	×
ReClor	5	×	√	√	×
CLUTRR	2	√	√	×	√
LTL	5	√	×	×	×
SoftReasoner	4	√	√	×	×
LogicNLI	7	√	√	×	×

Table 1: Comparisons among FOL datasets. Logic, Natural Language, and Commonsense correspond to three principles. #FOLs means how many FOLs are covered in the dataset, while Domain/Predicate indicates whether plenty of domain knowledge/predicate relations is/are required to solve the task.

ference. In addition, these two datasets do not provide proofs to trace back the reasoning process. CLUTRR (Sinha et al., 2019) also requires two FOLs but focuses more on the predicate relation (belongs to commonsense) understanding. LTL (Hahn et al., 2021) is a propositional logical benchmark containing five FOLs but does not conform to natural language. Clark et al. (2020) propose a series of novel FOL benchmarks (named SoftReasoner) that introduce as little commonsense as possible. It concentrates on a specific FOL combination, conjunctive implication with negation, rather than on diverse FOL forms. Inspired by SoftReasoner (Clark et al., 2020), we construct LogicNLI with common combinations of all seven FOLs to diagnose the FOL reasoning ability. Compared with other datasets (shown in Table 1), LogicNLI covers the most comprehensive FOL forms and effectively separates logic and commonsense. Furthermore, LogicNLI also provides all proofs for each instance so that we can evaluate LMs' FOL reasoning from different perspectives.

3 Task Definition

In this section, we introduce how the task on LogicNLI is defined. On the basis, we also exhibit how FOL reasoning is embodied in LogicNLI. We first define elements in LogicNLI: **Facts** $F = \{f_1, f_2, \cdots, f_n\}$ are composed of simple propositions; **Rules** $R = \{r_1, r_2, \cdots, r_m\}$ are always compound propositions with FOL; **Statement** s is the targeting proposition; **Premise** $P = (F, R)$ includes all facts and rules.

Based on the above definitions, the final objective of LogicNLI is to determine the logical relation between P and s under two assumptions: 1) World assumption is open (OWA); 2) The statement s and its negative expression $\neg s$ are independent conditioning on P ($\neg s \perp s|P$). The logical relations include "Entailment", "Contradiction", "Neutral", and "Paradox", whose conditions are shown in Equation 1, where $\vdash$ means syntactic consequence.

Facts: (F1) Harold is distinct. (F2) Daisy is not distinct. (F3) Alan is not distinct.

Rules: (R1) If someone is alive, then he is neither grieving nor worrisome. (R2) If there is at least one people who is distinct, then Alan is grieving. (R3) Harold being alive is equivalent to Alan being grieving. (R4) Someone being both worrisome and drab is equivalent to being colorful and distinct.

Statement: (S) Harold is grieving.
Proofs: (P1) Alan is grieving. (P2) Harold is alive.
Path: F1 + R2 → **P1** + R3 → **P2** + R1 → ¬S
Label: Contradiction

Figure 2: An instance in LogicNLI, including facts, rules, a statement, proofs, the path, and the label.

$$y = \begin{cases} Entailment, & P \vdash s \wedge P \nvdash \neg s \\ Contradiction, & P \nvdash s \wedge P \vdash \neg s \\ Neutral, & P \nvdash s \wedge P \nvdash \neg s \\ Paradox, & P \vdash s \wedge P \vdash \neg s \end{cases} \quad (1)$$

4 LogicNLI

4.1 Overview

LogicNLI includes more than 30K instances consisting of facts, rules, a statement to be judged, proofs, the reasoning path, and the label (shown in Figure 2). For each instance, it requires a multi-hop FOL reasoning process to reason out the final answer. To simplify the reasoning process, we set two limitations: 1) only considering the reasoning from cause to effect; 2) neglecting the true meanings of predicates. Therefore, LogicNLI is more suitable for benchmarking the specific (FOL) reasoning ability instead of serving as a comprehensive NLU task. As a result, we leave open the question of how LMs perform in real reasoning scenarios with FOLs because it is difficult to disentangle multiple influencing factors.

LogicNLI also provides four kinds of test sets that correspond to four diagnostic abilities in diagnosis, including total accuracy, robustness to irrelevant information, more-hop generalization, and proof-based traceability. Specifically, we attempt

to answer the following questions relevant to the FOL reasoning ability based on these evaluations: **Q1: Do models truly perform FOL reasoning automatically in diverse scenarios? Q2: Do reasoning results accord with reasonable logic?** Accuracy, robustness, and generalization are adopted to answer Q1 from different conditions. Accuracy is the most common in-domain evaluation that measures the overall performance of LMs. Compared with accuracy, the robustness test offers a scenario that increases/decreases non-proof sentences. As robustness does not change the reasoning process, it can be regarded as an in-domain evaluation. The generalization test offers a scenario that increases the reasoning hop and therefore increases proofs, so it is an out-of-domain evaluation. Traceability test is introduced to answer Q2 by validating the whole reasoning process according to the proofs.

4.2 Dataset Gerneration and Statistics

We adopt a semi-automatic method to generate LogicNLI with two steps: 1) logic generation, and 2) natural language generation. As for the logic generation, we adopt an automatic method to generate each logic expression to ensure the validity of FOL reasoning. Specifically, We first select a list of subjects, $S = \{s_i\}, i \leq n$, and a list of adjectives as predicates, $P = \{p_j\}, j \leq m$, and define a set of logical templates T in advance. For each instance, we randomly select logic expressions from T and the corresponding subjects and predicates from S and P. In terms of the natural language generation, we first adopt a rule-based method to generate initial language expressions and then make manual revisions. Manual correction aims to fix grammatical errors and semantic ambiguities. Besides, it also enhances the diversity of expressions. As for test sets of different abilities, we add additional limitations to generate data that meets different needs based on the above generation method.

Statistics of LogicNLI are listed in Table 2. LogicNLI includes 9 training sets, 9 development sets, and 15 test sets. We adopt different subjects and predicates for independently constructed training sets, development sets, and test sets to avoid the spurious correlations between subjects and predicates. To undermine the label bias, we ensure the balance of different labels in each dataset.

4.3 Diagnosis

Total Accuracy is the most intuitive indicator to measure the performance of a model in most NLU tasks (Storks et al., 2019), but it may not be sufficient as it cannot avoid the impacts of spurious correlations. In this work, the accuracy-test set (Test-A) has a similar distribution to the training set and the development set, except that the subjects and predicates are zero-shot.

Robustness to Irrelevant Information is an in-domain evaluation that measures the model's ability to extract relevant information from noisy data, which is typically the first step in many NLU tasks. Unlike Sinha et al. (2019), our work focuses on the amount of noise, rather than its taxonomy. Therefore, we adopt an elimination method to generate training sets (Train-R), development sets (Dev-R), and test sets (Test-R). Firstly, facts and rules are classified into relevant sentences (R1, R2, and R3 in Figure 2) and irrelevant sentences (R4 in Figure 2). Secondly, we fix the relevant sentences to ensure that the label remains unchanged and gradually eliminate irrelevant ones. We finally acquire robustness sets with different numbers of facts and rules (from 10 to 24 in steps of 2).

More-hop Generalization is an out-of-domain indicator to judge whether a model truly understands the logic rules and applies them to reasoning instances. Following the setting in CLUTRR (Sinha et al., 2019), generalization can be measured by training a model on examples with $\leq$k-hop reasoning and evaluated on ones with $>$k-hop reasoning. Therefore, we generate a series of the more-hop test sets (Test-G) only by controlling the generation iterations during the logic generation.

Proof-based Traceability is used to post-verify whether a model infers the correct answer according to the human-understandable logic. In multi-hop reasoning tasks, it is reasonable to measure traceability through proofs (Yang et al., 2018; Gontier et al., 2020). Therefore, we propose proof-based traceability (the example of proofs is shown in Figure 2) based on the intuitive that if a model can infer the correct answer according to the right reasoning paths, it will correctly validate each proof. Specifically, we construct an traceability-test set (Test-T) with 6-hop instances to make the final task an out-of-domain evaluation while ensuring the judgments of proofs are in-domain. Since "Neutral" samples do not provide any proofs, we remove them. To perform the diagnosis, we first train the model on the training set and test it on Test-T.

Data	Statistics	Train	Dev.	Test-A	Train-R(s)	Dev-R(s)	Test-R(s)	Test-G(s)	Test-T
d-LogicNLI	#Instances	12000	1500	1500	72000	9000	9000	9396	6094
	Avg. Length	184	182	183	63/ 87/ 111/ 136/ 160/ 184*			342	340
	Max. Length	215	212	212	133/ 140/ 167/ 195/ 206/ 232*			389	391
	#Hop	≤5	≤5	≤5	≤5			6/ 7/ 8/ 9/ 10	6
	#(Facts+Rules)	15	15	15	5/ 7/ 9/ 11/ 13/ 15			15	15
	#Subjects(n)	382	100	100	382	100	100	100	100
	#Predicates(m)	379	100	100	379	100	100	100	100
	%Labels	Entailment: Contradiction: Neutral = 1: 1: 1 (1:1:0 for Test-T)							
LogicNLI	#Instances	16000	2000	2000	96000	16000	16000	4124	6039
	Avg. Length	245	245	245	104/ 125/ 145/ 165/ 185/ 205/ 225/ 245*			330	339
	Max. Length	291	279	272	167/ 188/ 227/ 230/ 250/ 274/ 283/ 291*			395	401
	#Hop	≤5	≤5	≤5	≤5			6/ 7/ 8/ 9/ 10	6
	#(Facts+Rules)	24	24	24	10/ 12/ 14/ 16/ 18/ 20/ 22/ 24			24	24
	#Subjects(n)	382	100	100	382	100	100	100	100
	#Predicates(m)	379	100	100	379	100	100	100	100
	%Labels	Entailment: Contradiction: Neutral: Paradox = 1: 1: 1: 1 (1:1:0:1 for Test-T)							

Table 2: Statistical information for d-LogicNLI and LogicNLI datasets. A, R, G, and T represent accuracy, robustness, generalization, and traceability, respectively. *Length information under different #(Rules+Facts) is provided as the length distributions are different in robustness sets.

Next, we extract the instances that are correctly predicted to form the target set. We then revise all proofs of the target set to positive expressions to avoid the "negation" logic's impact on the evaluation and re-annotate them. Finally, inspired by the exact match metric (Yang et al., 2018), we define a proof-based extract match (P-EM) to calculate the percentage of instances whose proofs are completely correctly predicted. We adopt P-EM and proof accuracy (P-Acc) to measure the traceability.

4.4 Degraded LogicNLI

"Paradox" provides a virtual scenario that is not common in texts, so most classic NLI tasks do not have this condition. To further understand why we introduce "Paradox" to LogicNLI, we construct a degraded dataset, named d-LogicNLI, as a comparison. Compared with LogicNLI, d-LogicNLI only contains premises and hypotheses with logical relations of "Entailment", "Contradiction", and "Neutral". From the perspective of dataset construction, we only need to set a filter in the logic generation stage to filter out paradox propositions. The statistics of d-LogicNLI are listed in Table 2.

5 Experiments

5.1 Experimental Settings

We conduct experiments on three state-of-the-art language models (LMs), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), to systematically measure their FOL reasoning ability. For a fair comparison, we fine-tune the large versions of LMs with the same hidden size (1024) and adopt a two-layer perceptron to predict the logical relation. Following the input form of NLI tasks, the inputs look like "[CLS] facts rules [SEP] statement [SEP]" for BERT and RoBERTa, and "facts rules [SEP] statement [SEP] [CLS]" for XLNet. The hyper-parameters are shown in Table 3. We set random selection and human performance as the lower and upper boundaries of accuracy. As for human performance evaluation, we employ four Ph.D. students and five post-graduate students of different majors, reporting the average scores on 500 randomly selected instances from the development and test sets. We consider a question as being correctly answered if one of the students gives the correct answer.

Paras.	**BERT**	**RoBERTa**	**XLNET**
batch size	16	16	16
lr	–	$1e^{-5}$	$1e^{-3}$
lr for BERT	$5e^{-6}$	$5e^{-6}$	–
decay rate	0.9	0.9	0.8
l2 coeff.	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$
early stop	5	5	5
epochs	20	20	20
optimizer	ADAMW	ADAMW	ADAMW

Table 3: Hyper-parameter settings.

5.2 Results

Total Accuracy. From Table 4, all three LMs perform better than random guess (25.0%) but worse than humans (77.5%). RoBERTa performs the best

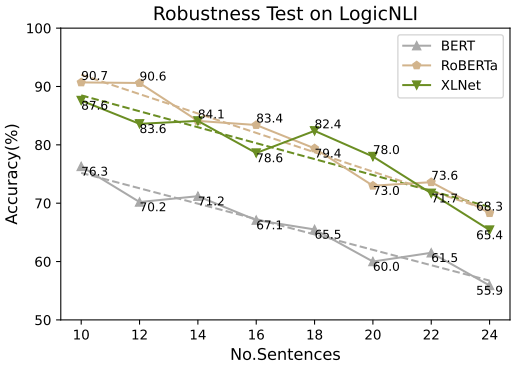

Figure 3: Robustness analysis on LogicNLI. All LMs are trained on Train-R with different number of sentences and tested on the Test-R. Line graphs show changes of accuracies with increasing number of sentences. Dashed lines are linear fitting equations.

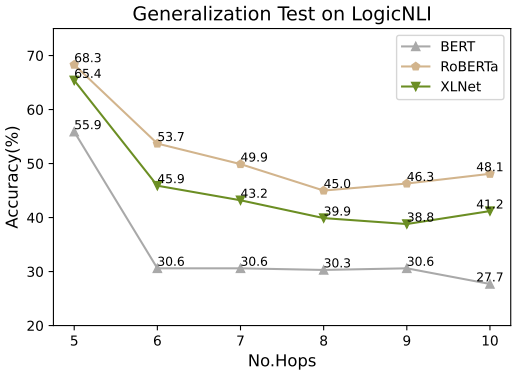

Figure 4: Generalization analysis based on LogicNLI. Results on 6, 7, 8, 9, and 10-hop sets are regarded as out-of-domain results, while it on ≤5-hop set is an in-domain result.

on both the development dataset (65.0%) and Test-A (68.3%), with a gap of fewer than ten points compared with humans on Test-A. XLNet is slightly inferior to RoBERTa with the accuracies of 64.0% and 65.4% on the development dataset and Test-A, respectively. BERT, the worst LMs of the three, only achieves only 57.0% and 55.9% accuracies on two datasets, which are significantly poorer than humans. Overall, from the perspective of accuracy, all three LMs cannot reach the human level.

Robustness to Irrelevant Information. Table 4 shows the average results on all Dev-R(s) and Test-R(s). Similar to accuracy, RoBERTa's performance is slightly better than XLNet, but the gap between the two is not significant. BERT still performs the worst on both Dev-R and Test-R.

Average accuracy on Test-R(s) cannot effectively reflect the robustness directly. We plot the line graph that describes the trend of the result on Test-R(s) with the change of the number of sentences (facts+rules) in Figure 3. All three LMs show downward trends as the number of irrelevant sentences increases. The performances of BERT and RoBERTa decrease evenly with the noise increasing, while the performance of XLNet is fluctuating in the former period but declines rapidly in the latter. Furthermore, we calculate the degradation rate δ_R from the 10-sentence Test-R to 24-sentence Test-R to measure robustness. Since the descent process is non-linear, we replace original polylines with their fitting lines (dotted lines in Figure 3) to ensure that the degradation rate includes all test points' information. The final degradation rates of BERT, RoBERTa, and XLNet are 24.6%, 25.2%, and 21.5%, which shows that XLNet's robustness is slightly better than BERT and RoBERTa.

More-hop Generalization. We plot accuracies on Test-A and each Test-G in Figure 4 and show the total accuracy on Test-G in Table 4. From Figure 4, all three LMs' performances have dramatically dropped when transferring from in-domain scenarios to out-of-domain scenarios. However, their out-of-domain accuracies can almost keep stable as the number of hops continues to increase (up to 10). To further compare the generalization, we define an indicator, $\delta_{A\to G} = \frac{M_1 - M_2}{M_1} \times 100\%$, to reflect the percentage of performance degradation when transferring from in-domain scenarios to out-of-domain scenarios, where M_1 is the in-domain result on Test-A and M_2 is the average out-of-domain result on Test-G. The performance degradation rates of BERT, RoBERTa, and XLNet are 43.5%, 26.9%, and 34.3%, respectively. Therefore, RoBERTa shows the best generalization when transferring to more-hop reasoning, while BERT cannot effectively understand logical rules and apply them to out-of-domain instances.

Proof-based Traceability. Considering P-Acc on Test-T (Table 4), it seems that 87.6% of proofs can be validated when adopting RoBERTa to make the prediction. Even BERT can explain more than 60% proofs. However, we usually judge whether an instance is understood logically by verifying the completeness of the whole logical chain instead of the ratio of understandable proofs. Therefore, P-EM is more suitable than P-Acc to measure traceability. Considering EM, RoBERTa can validate 53.1% correctly predicted instances, while BERT and XLNet can only validate 9.3% and

Models	Accuracy		Robustness		Generalization	traceability			
	Dev.	Test-A	Dev-R	Test-R	Test-G	#Target	Test-T(P-EM)	#Proof	Test-T(P-Acc.)
Random	25.0		25.0		25.0		-		
Human	77.5		-		-		-		
BERT	57.0	55.9	68.0	66.0	31.6	2143	9.3	12706	61.1
RoBERTa	**65.0**	**68.3**	**80.9**	**80.4**	**49.9**	3529	**53.1**	21728	**87.6**
XLNet	64.0	65.4	77.0	78.9	43.0	2495	28.6	15112	77.0

Table 4: Diagnostic results of LMs on LogicNLI. All information provide the percentage (%) of each evaluation except for #Target and #Proof.

28.6% instances, respectively. This result means that RoBERTa is the only LM that can perform FOL reasoning to some extent, which has significantly better proof-based traceability than BERT and XLNet. However, even the best model, RoBERTa, can only explain approximately half of the predictions, indicating that the overall predictions made by LMs do not conform to human logic.

5.3 Overall Diagnosis

Considering four evaluations comprehensively, RoBERTa has the best FOL reasoning ability in complex scenarios and is the only one of three LMs that can provide a certain degree of traceability. Considering accuracy (in-domain evaluation) and generalization (out-of-domain evaluation), RoBERTa performs significantly better than BERT and XLNet. Especially when transferring from the in-domain scenarios to the out-of-domain, RoBERTa's degradation ratio is significantly lower than BERT's and XLNet's, which means that RoBERTa is better at understanding logical rules and applying them than the other two LMs. This conclusion can also be proven by the traceability test. In reality, although BERT and XLNet can make correct predictions to some extent, most of these results cannot be traced back by the validation of proofs. A certain percentage of prediction results of RoBERTa can be explained. Finally, as for robustness, RoBERTa is indeed more susceptible to irrelevant information than XLNet. Even though, RoBERTa still performs better than XLNet on robustness test, as XLNet's performance drops rapidly after reaching a threshold.

In general, even for RoBERTa, there is still a long way to the real FOL reasoning. On the one hand, its performance needs to be improved in both in-domain and out-of-domain scenarios. On the other hand, even if RoBERTa makes the correct prediction, nearly half of its prediction results are still unexplainable. The gap between LMs and humans motivates us to explore more effective ways to make more effective reasoning in NLU. Maybe neural symbolic models are solutions to FOL reasoning (Kalouli et al., 2020).

Statistics	$\wedge$	$\vee$	$\neg$	$\rightarrow$	$\equiv$	$\forall$	$\exists$
#Instances	1500	1500	1500	1500	1500	1500	1500
Ave.Length	358	330	270	246	253	275	292
Max.Length	411	364	299	292	279	333	321
#(Facts+Rules)	25	25	25	24	24	30	22
#Subjects(n)	100	100	100	100	100	100	100
#Predicates(m)	100	100	100	100	100	100	100
%Labels	Entailment: Contradiction: Neutral = 1: 1: 1						

Table 5: Statistical information for each FOL.

Models	Performance on Each FOL						
	$\wedge$	$\vee$	$\neg$	$\rightarrow$	$\equiv$	$\forall$	$\exists$
Random	33.3	33.3	33.3	33.3	33.3	33.3	33.3
Human	88.9	100	100	93.3	90.0	96.0	88.0
BERT	58.9	65.8	62.8	68.2	64.9	66.8	76.9
RoBERTa	**82.1**	**94.7**	**68.5**	**87.1**	**84.4**	**80.5**	**99.3**
XLNet	78.0	90.6	66.2	81.2	80.1	75.0	98.3

Table 6: Performance on each FOLs (%). Except for $\neg$, other FOLs cannot imply "Paradox", so we remove "Paradox" and the random accuracy is 33.3%.

5.4 Analysis of Each FOLs

To further understand the FOL reasoning ability, we perform the analysis on how LMs understand each FOL. Specifically, we are required to disentangle the target FOL from other FOLs by adding logical filters when selecting filters. Among seven FOLs, only implication and equivalence can be fully entangled from other FOLs and directly used for reasoning, while others alone cannot constitute complete reasoning. Therefore, we combine the other five FOLs with the implication logic to make the reasoning process effective. Statistics are shown in Table 5.

Results of FOLs experiments are shown in Ta-

ble 6. RoBERTa outperforms the other two LMs on all FOLs. Considering each FOL, the performance of LMs is almost difficult to surpass humans, except on the existential logic. In reality, existential logic is difficult for humans (with the lowest human performance) because it requires traversing all information to extract relevant information. However, it is not difficult for LMs as existential logic provides weak constraints that are easy to satisfy. As a result, most LMs perform better than humans on such logic. On the contrary, LMs' performances on universal logic and negation logic are significantly worse than humans'. As for universal logic, its complexity may come from its ambiguity in language. For example, comparing $\forall x F(x) \rightarrow G(a)$ and $\forall x(F(x) \rightarrow G(x))$, although both use universal logic for reasoning, the former requires stronger conditions but can only provide simpler conclusions than the latter. This phenomenon makes universal logic difficult to understand consistently. In terms of negation, many studies (Hossain et al., 2020b,c,a) have proved that negation logic itself is critical but difficult to be understood by neural networks, which results in more auxiliary methods to identify and process in natural language. In addition, we find that all LMs perform better on single FOL datasets than on Test-A, which is evidence that LMs suffer from the coupling of different FOLs. Therefore, the analysis of FOLs motivates us to modify LMs by 1) focusing on specific logic types (negation and universal logic), and 2) disentangling the different logical forms.

5.5 Analysis of "Paradox"

In this section, we provide further analysis on why to introduce the virtual label "Paradox" into LogicNLI by comparing d-LogicNLI and LogicNLI. As shown in Figure 5, d-LogicNLI is a particular case of LogicNLI under the mutually exclusive condition of "Entailment" and "Contradiction". Therefore, although "Paradox" is usually a virtual label in most scenarios, it is critical to complement the space of the logical relation.

In practice, we can summarize two effects of "Paradox": 1) "Paradox" provides more accurate FOL information for model training, thereby effectively suppressing the impacts of spurious correlations caused by dataset bias; 2) "Paradox" makes the diagnostic scenarios more complete and complex, so it can better distinguish the FOL reasoning abilities of different LMs. We will illustrate

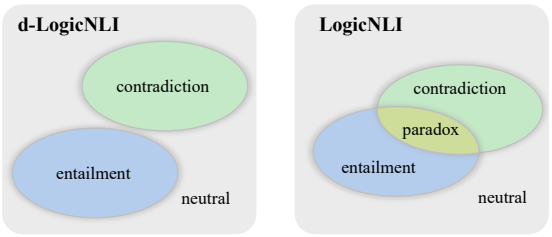

Figure 5: Comparison of logical relation spaces of d-LogicNLI and LogicNLI.

Models	Data	P-Acc	δ_R	$\delta_{A\rightarrow G}$	P-EM
BERT	d-LogicNLI	**73.1**	**23.7**	44.0	1.1
	LogicNLI	55.9	24.6	**43.5**	**9.3**
RoBERTa	d-LogicNLI	**80.7**	**17.0**	40.0	0.9
	LogicNLI	68.3	25.2	**26.9**	**53.1**
XLNet	d-LogicNLI	**85.7**	**7.2**	36.8	4.1
	LogicNLI	65.4	21.5	**34.3**	**28.6**

Table 7: Comparison of important indicators on d-LogicNLI and LogicNLI. δ_R and $\delta_{A\rightarrow G}$ are degradation rates introduced in the results of robustness and generalization, respectively. P-EM is a metric to measure traceability.

these two statements by comparing important indicators on d-LogicNLI and LogicNLI (shown in Table 7). Firstly, the in-domain results (Accuracy and δ_R) of all three LMs (and human performance) on d-LogicNLI are overall better than those on LogicNLI, proving that either d-LogicNLI provides much simpler evaluation datasets than LogicNLI does, or d-LogicNLI provides more precise and unbiased training instances than LogicNLI provides. Secondly, we observe that LMs trained on d-LogicNLI are hardly traceable based on Test-T (the maximum P-EM achieved by XLNet is only 4.1%), while LMs trained on LogicNLI have significantly better traceability. This phenomenon support that d-LogicNLI does not provide sufficient information for LMs to master the FOL reasoning ability. Finally, the generalization indicators $\delta_{A\rightarrow G}$ of BERT, RoBERTa, and XLNet trained on d-LogicNLI are 44.0%, 40.0%, and 36.8%, respectively, showing that the transferring ability of LMs trained on d-LogicNLI is not as good as those trained on LogicNLI. This is implicit evidence to support that LogicNLI provides more information for LMs to understand FOL rules.

5.6 Discussion

From Table 4, RoBERTa performs the best on LogicNLI while XLNet outperforms the other two LMs

on d-LogicNLI. According to the original work of these LMs (Liu et al., 2019; Yang et al., 2019), XLNet modifies the architecture of BERT, while RoBERTa mainly introduces a larger corpus to train the model. In most simple reasoning scenarios, such as RACE (Lai et al., 2017) and SQuAD (Rajpurkar et al., 2016), the performance of XLNet is usually better than RoBERTa's. However, in other scenarios that require more complicated reasoning processes, such as LogiQA (Liu et al., 2020) and datasets defined in GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a), RoBERTa, trained on a larger corpus, usually outperforms XLNet. Based on the above analysis, LogicNLI provides more complex reasoning scenarios than d-LogicNLI. Therefore, RoBERTa can highlight its advantages even more on LogicNLI.

6 Conclusion

In this paper, we propose a diagnostic method to diagnose LMs' FOL reasoning ability. This method introduces a novel proposed benchmark, LogicNLI, that disentangles the FOL reasoning from commonsense inference. Specifically, it includes four evaluations to measure the FOL reasoning ability from different perspectives. Results on three LMs show that although some LMs (RoBERTa) own a certain interpretable FOL reasoning ability, they still cannot make sensible FOL reasoning like humans. Detailed analysis motivates us to enhance specific reasoning abilities or explore new methods to make neural models understand more refined logic.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2018YFC0830400), the ECNU-SJTU joint grant from the Basic Research Project of Shanghai Science and Technology Commission (19JC1410102), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Shanghai Science and Technology Innovation Action Plan (20511102600).

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL-HLT*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *ICLR*.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *AAAI*, pages 7432–7439.

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

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *IJCAI*.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *EMNLP-IJCNLP*.

Ernest Davis. 2017. Logical formalizations of commonsense reasoning: A survey. *JAIR*, 59:651–723.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT*.

Nicolas Gontier, Koustuv Sinha, Siva Reddy, and Christopher Pal. 2020. Measuring systematic generalization in neural proof generation with transformers. In *NeurIPS*.

Christopher Hahn, Frederik Schmitt, Jens U. Kreber, Markus Norman Rabe, and Bernd Finkbeiner. 2021. Teaching temporal logics to neural networks. In *ICLR*.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020a. It's not a non-issue: Negation as a source of error in machine translation. In *EMNLP(Findings)*.

Md Mosharaf Hossain, Kathleen E. Hamilton, Alexis Palmer, and Eduardo Blanco. 2020b. Predicting the focus of negation: Model and error analysis. In *ACL*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020c. An analysis of natural language inference benchmarks through the lens of negation. In *EMNLP*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *EMNLP-IJCNLP*, pages 2391–2401.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. In *ACL*.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *ACL*.

Aikaterini-Lida Kalouli, Richard S. Crouch, and Valeria de Paiva. 2020. Hy-nli: a hybrid system for natural language inference. In *COLING*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *ACL*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *IJCAI*.

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*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Penstein Rosé, and Eduard H. Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *CoNLL*.

Marco Túlio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *ACL*.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *CoRR*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *ACL*.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *CoRR*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *AAAI*.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. Quartz: An open-domain dataset of qualitative relationship questions. In *EMNLP-IJCNLP*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics - on what language model pre-training captures. *TACL*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *ICLR*.