IDOL: Indicator-oriented Logic Pre-training for Logical Reasoning

Zihang Xu[†], **Ziqing Yang**[†], **Yiming Cui**[‡†], **Shijin Wang**[†§]
[†]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[‡]Research Center for SCIR, Harbin Institute of Technology, Harbin, China
[§]iFLYTEK AI Research (Central China), Wuhan, China
[†]{zhxu13,zqyang5,ymcui,sjwang3}@iflytek.com
[‡]ymcui@ir.hit.edu.cn

Abstract

In the field of machine reading comprehension (MRC), existing systems have surpassed the average performance of human beings in many tasks like SQuAD. However, there is still a long way to go when it comes to logical reasoning. Although some methods for it have been put forward, they either are designed in a quite complicated way or rely too much on external structures. In this paper, we proposed **IDOL** (**I**n**D**icator-**O**riented **L**ogic Pre-training), an easy-to-understand but highly effective further pre-training task which logically strengthens the pre-trained models with the help of 6 types of logical indicators and a logically rich dataset **Lo**ng**G**ic **P**re-training (LGP). IDOL achieves state-of-the-art performance on ReClor and LogiQA, the two most representative benchmarks in logical reasoning MRC, and is proven to be capable of generalizing to different pre-trained models and other types of MRC benchmarks like RACE and SQuAD 2.0 while keeping competitive general language understanding ability through testing on tasks in GLUE. Besides, at the beginning of the era of large language models, we take several of them like ChatGPT into comparison and find that IDOL still shows its advantage.[1]

1 Introduction

With the development of pre-trained language models, a large number of tasks in the field of natural language understanding have been dealt with quite well. However, those tasks emphasize more on assessing basic abilities like word-pattern recognition of the models while caring less about advanced abilities like reasoning over texts (Helwe et al., 2021).

In recent years, an increasing number of challenging tasks have been brought forward gradually. At sentence-level reasoning, there is a great variety of benchmarks for natural language inference like QNLI (Demszky et al., 2018) and MNLI (Williams et al., 2018). Although the construction processes are different, nearly all these datasets evaluate models with binary or three-way classification tasks which need reasoning based on two sentences. At passage-level reasoning, the most difficult benchmarks are generally recognized as the ones related to logical reasoning MRC which requires question-answering systems to fully understand the whole passage, extract information related to the question and reason among different text spans to generate new conclusions in the logical aspect. In this area, the most representative benchmarks are some machine reading comprehension datasets like ReClor (Yu et al., 2020) and LogiQA (Liu et al., 2020).

Considering that there are quite few optimization strategies for the pre-training stage and that it is difficult for other researchers to follow and extend the existing methods which are designed in rather complex ways, we propose an easy-to-understand but highly effective pre-training task named IDOL which helps to strengthen the pre-trained models in terms of logical reasoning. We apply it with our customized dataset LGP which is full of logical information. Moreover, we experimented with various pre-trained models and plenty of different downstream tasks and proved that IDOL is competitive while keeping models and tasks agnostic.

Recently, ChatGPT attracts a lot of attention all over the world due to its amazing performance in question answering. Thus, we also arranged an experiment to let IDOL compete with a series of LLMs (large language models) including it.

The contributions of this paper are summarized as follows:

- Put forward the definitions of 5 different types of logical indicators. Based on these we construct the dataset LGP for logical pre-training and we probe the impact of different types of logical indicators through a series of ablation experiments.

[1]Please refer to https://github.com/GeekDream-x/IDOL for relevant resources including datasets, models, and codes.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8099–8111
July 9-14, 2023

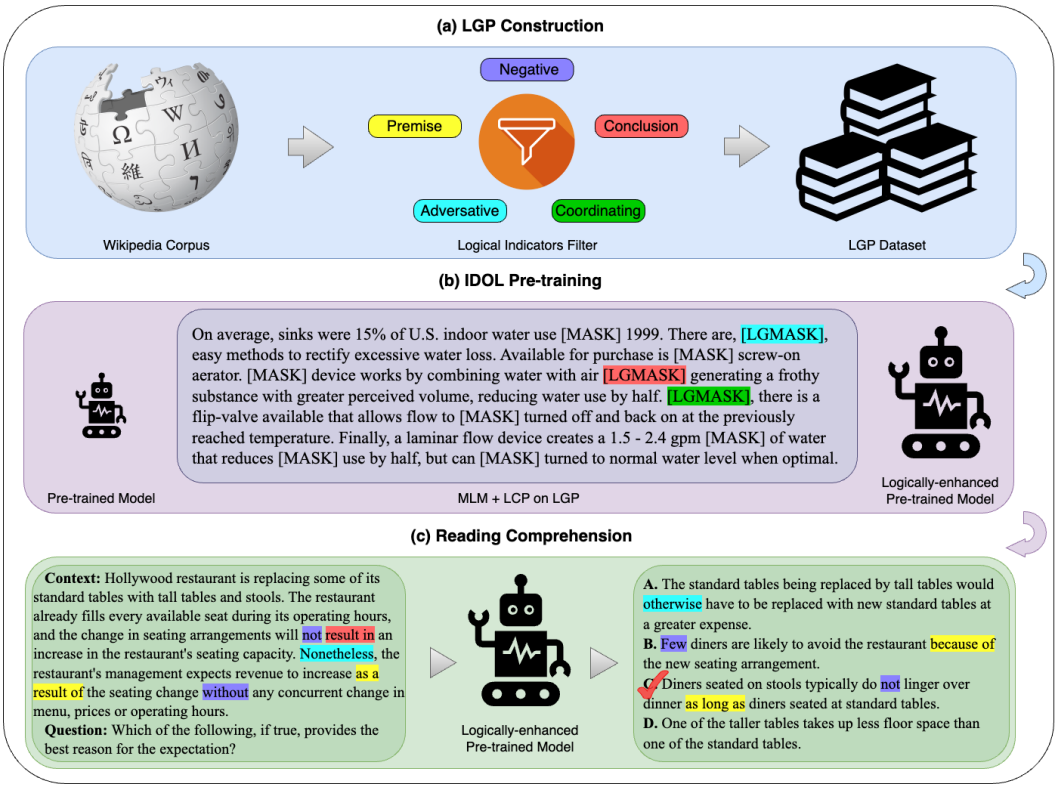

Figure 1: A diagram illustrating the three steps of our method: (a) construct the logically rich dataset LGP from Wikipedia, (b) further pre-train models to improve logical reasoning ability, and (c) answer logical reasoning MRC questions with the help of logical indicators appeared both in context and choices. See Section 4 for more details on our method.

- Design an indicator-oriented further pre-training method named IDOL, which aims to enhance the logical reasoning ability of pre-trained models. It achieves state-of-the-art performance in logical reasoning MRC and shows progress in general MRC and general understanding ability evaluation.

- The first to provide a pilot test about the comparison between fine-tuning traditional pre-trained models and prompting LLMs in the field of logical reasoning MRC.

2 Related Work

2.1 Logical Reasoning

In order to help reasoning systems perform better on reading comprehension tasks focusing on logical reasoning, there have been a great many methods put forward by research institutions from all over the world. Unsurprisingly, the majority of the optimization approaches put forward revolve around the fine-tuning phase while there are far fewer methods designed for further pre-training.

In the aspect of pre-training, to the best of our knowledge, there are only two approaches presented in published papers called MERIt and LogiGAN. MERIt team generated a dataset from the one provided by Qin et al. (2021) which contains passages from Wikipedia with annotations about entities and relations. And then optimize the model on that with the help of contrastive learning (Jiao et al., 2022). The researchers behind LogiGAN use a task about statement recovery to enhance the logic understanding ability of generative pre-trained language models like T5 (Pi et al., 2022).

For optimizing models at the fine-tuning phase, there are dozens of methods proposed as far as we know. For example, LReasoner put forward a context extension framework with the help of logical equivalence laws including contraposition and transitive laws (Wang et al., 2022a). Another example is Logiformer which introduced a two-stream architecture containing a syntax branch and a logical branch to better model the relationships among distant logical units (Xu et al., 2022).

Type	Library	Example
PMI	given that, seeing that, for the reason that, owing to, as indicated by, on the grounds that, on account of, considering, because of, due to, now that, may be inferred from, by virtue of, in view of, for the sake of, thanks to, as long as, based on that, as a result of, considering that, inasmuch as, if and only if, according to, in that, only if, because, depend on, rely on	The real world contains no political entity exercising literally total control over even one such aspect. This is because any system of control is inefficient, and, therefore, its degree of control is partial.
CLI	conclude that, entail that, infer that, that is why, therefore, thereby, wherefore, accordingly, hence, thus, consequently, whence, so that, it follows that, imply that, as a result, suggest that, prove that, as a conclusion, conclusively, for this reason, as a consequence, on that account, in conclusion, to that end, because of this, that being so, ergo, in this way, in this manner, by such means, as it turns out, result in, in order that, show that, eventually	In the United States, each bushel of corn produced might result in the loss of as much as two bushels of topsoil. Moreover, in the last 100 years, the topsoil in many states, which once was about fourteen inches thick, has been eroded to only six or eight inches.
NTI	not, neither, none of, unable, few, little, hardly, merely, seldom, without, never, nobody, nothing, nowhere, rarely, scarcely, barely, no longer, isn't, aren't, wasn't, weren't, can't, cannot, couldn't, won't, wouldn't, don't, doesn't, didn't, haven't, hasn't	A high degree of creativity and a high level of artistic skill are seldom combined in the creation of a work of art.
ATI	although, though, but, nevertheless, however, instead of, nonetheless, yet, rather, whereas, otherwise, conversely, on the contrary, even, nevertheless, despite, in spite of, in contrast, even if, even though, unless, regardless of, reckless of	This advantage accruing to the sentinel does not mean that its watchful behavior is entirely self-interested. On the contrary, the sentinel's behavior is an example of animal behavior motivated at least in part by altruism.
CNI	and, or, nor, also, moreover, in addition, on the other hand, meanwhile, further, afterward, next, besides, additionally, meantime, furthermore, as well, simultaneously, either, both, similarly, likewise	A graduate degree in policymaking is necessary to serve in the presidential cabinet. In addition, everyone in the cabinet must pass a security clearance.

Table 1: Libraries and examples of all types of logical indicators.

2.2 Pre-training Tasks

As NLP enters the era of pre-training, more and more researchers are diving into the design of pre-training tasks, especially about different masking strategies. For instance, in Cui et al. (2020), the authors apply Whole Word Masking (WWM) on Chinese BERT and achieved great progress. WWM changes the masking strategy in the original masked language modeling (MLM) into masking all the tokens which constitute a word with complete meaning instead of just one single token. In addition, Lample and Conneau (2019) extends MLM to parallel data as Translation Language Modeling (TLM) which randomly masks tokens in both source and target sentences in different languages simultaneously. The results show that TLM is beneficial to improve the alignment among different languages.

3 Preliminary

3.1 Text Logical Unit

It is admitted that a single word is the most basic unit of a piece of text but its meaning varies with different contexts. In Xu et al. (2022), the authors refer logical units to the split sentence spans that contain independent and complete semantics. In this paper, since much more abundant logical indicators with different types that link not only clauses but also more fine-grained text spans are introduced, we extend this definition to those shorter text pieces like entities.

3.2 Logical Indicators

By analyzing the passages in logical reasoning MRC and reasoning-related materials like debate scripts, we found that the relations between logic units (like entities or events) can be summarized into 5 main categories as follows and all these relations are usually expressed via a series of logical indicators. After consulting some previous work like Pi et al. (2022) and Penn Discourse TreeBank 2.0 (PDTB 2.0) (Prasad et al., 2008), we managed to construct an indicator library for each category. As for the examples of indicators we used in detail, please refer to Table 1.

- **Premise/Conclusion Indicator (PMI/CLI)** The first two types of logical indicators pertain to premises and conclusions. These indicators

signal the logical relationship between statements. For instance, premise expressions such as "due to" indicate that the logic unit following the keyword serves as the reason or explanation for the unit preceding it. Conversely, conclusion phrases like "result in" suggest an inverse relationship, implying that the logic unit after the keyword is a consequence or outcome of the preceding unit.

- **Negative Indicator (NTI)** Negative indicators, such as "no longer", play a crucial role in text logic by negating affirmative logic units. They have the power to significantly alter the meaning of a statement. For example, consider the sentences "Tom likes hamburgers." and "Tom no longer likes hamburgers." These two sentences have nearly opposite meanings, solely due to the presence of the indicator "no longer".

- **Adversative Indicator (ATI)** Certain expressions, such as "however", are commonly employed between sentences to signify a shift or change in the narrative. They serve as valuable tools for indicating the alteration or consequence of a preceding event, which helps to cover this frequent kind of relation among logic units.

- **Coordinating Indicator (CNI)** The coordinating relation is undoubtedly the most prevalent type of relationship between any two logic units. Coordinating indicators are used to convey that the units surrounding them possess the same logical status or hold equal importance. These indicators effectively demonstrate the coordination or parallelism between the connected logic units.

4 Methodology

4.1 LGP Dataset Construction

For the sake of further pre-training models with IDOL, we constructed the dataset LGP (LoGic Pre-training) based on the most popular unannotated corpus English Wikipedia.[2] We first split the articles into paragraphs and abandoned those whose lengths (after tokenization) were no longer than 5. In order to provide as much logical information as possible, we used the logical indicators listed in Table 1 to filter the Wiki paragraphs. During this procedure, we temporarily removed those indicators with extremely high frequency like "and", otherwise, there would be too many paragraphs whose logical density was unacceptably low. Then, we iterated every logical keyword and replaced it with our customized special token `[LGMASK]` under the probability of 70%.

For the purpose of modeling the ability to distinguish whether a certain masked place is logic-related or not, we introduced the sixth logical indicator type - Logic Unrelated Indicator (LUI). Based on this, we then randomly replaced 0.6% tokens other than logical indicators with `[LGMASK]`. Afterward, the labels for the logical category prediction (LCP) task were generated based on the corresponding logic types of all the `[LGMASK]`s. In the end, take RoBERTa (Liu et al., 2019) for example, our logic dataset LGP contains over 6.1 million samples and as for the quantities of logical indicators in each type please refer to Figure 2.

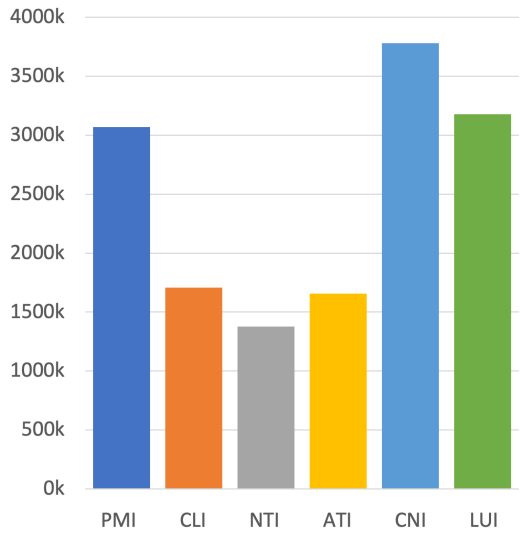

Figure 2: The numbers of 6 types of logical indicators in LGP for RoBERTa.

4.2 IDOL Pre-training

4.2.1 Logical Category Prediction

As introduced in section 3.2 and section 4.1, we defined a logic-related special mask token `[LGMASK]` and it will take the place of 6 types of logical indicators - PMI, CLI, NTI, ATI, CNI, and LUI. During the forward process of fine-tuning the pre-trained models, the corresponding logical categories need to be predicted by them like what will be done in the token classification task of the standard Masked Language Modeling (MLM) (Devlin et al., 2019).

[2] https://dumps.wikimedia.org/

When the models are trying to predict the correct logical type of a certain [LGMASK], they will learn to analyze the relationship among the logical units around the current special token and whether there is some kind of logical relations with the help of the whole context. Therefore, the pre-trained models will be equipped with a stronger ability of reasoning over texts gradually.

Moreover, we use Cross-Entropy Loss (CELoss) to evaluate the performance of predicting the logical categories. The loss function for LCP is as described in Equation (1) where n is the number of samples, m is the number of [LGMASK] in the i_{th} sample, $y_{i,j}$ indicates the model prediction result for the j_{th} [LGMASK] in the i_{th} sample and $\hat{y}_{i,j}$ denote the corresponding ground truth value.

$$\mathcal{L}_{\text{LCP}} = \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \text{CELoss}(y_{i,j}, \hat{y}_{i,j}) \quad (1)$$

4.2.2 IDOL

To avoid catastrophic forgetting, we combine the classic MLM task with the LCP introduced above to become IDOL, a multi-task learning pre-training method for enhancing the logical reasoning ability of pre-trained models. For the purpose of balancing the effects of the two pre-training tasks, we introduced a hyper-parameter λ as the weight of the loss of LCP (the proper λ depends on the pre-trained language model used and the empirical range is between 0.7 and 0.9). Thus, for the IDOL pre-training loss function, please refer to Equation (2). Figure 3 presented an example of IDOL pre-training where predicting tokens and the classes of logical indicators simultaneously.

$$\mathcal{L}_{\text{IDOL}} = \lambda \cdot \mathcal{L}_{\text{LCP}} + (1 - \lambda) \cdot \mathcal{L}_{\text{MLM}} \quad (2)$$

5 Experiments

5.1 Baselines

With the rapid development of pre-training technology these years, we have various choices for backbone models. In this paper, we decide to apply IDOL on BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), ALBERT-xxlarge (Lan et al., 2020) and DeBERTa-v2-xxlarge (He et al., 2021) and will evaluate the models in the following three different aspects in section 5.4 to better verify the performance of IDOL.[3]

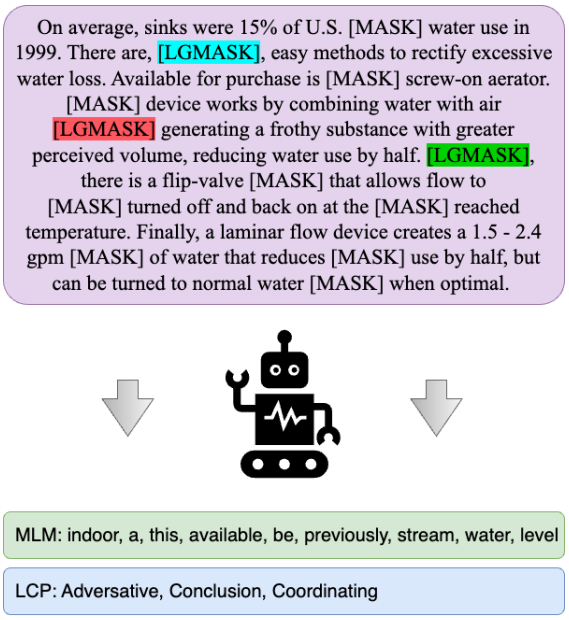

Figure 3: An example of pre-training with IDOL. The model needs to recover the tokens replaced by [MASK] (MLM) and predict the category of each logical indicator masked by [LGMASK] (LCP) in the meantime.

In terms of logical reasoning MRC, we will compare IDOL with several previous but still competitive methods for logical reasoning MRC including DAGN (Huang et al., 2021), AdaLoGN (Li et al., 2022), LReasoner (Wang et al., 2022b), Logiformer (Xu et al., 2022) and MERIt (Jiao et al., 2022). Much more interesting, we let IDOL compete with ChatGPT in a small setting.

5.2 Datasets

First and foremost, the aim of IDOL is to improve the logical reasoning ability of pre-trained models, thus, the two most representative benchmarks - ReClor and LogiQA will act as the primary examiners.

Following this, RACE (Lai et al., 2017) and SQuAD 2.0 (Rajpurkar et al., 2018), two classic machine reading comprehension datasets that are not targeted at assessing reasoning ability, will come on stage, which will be beneficial to conclude whether IDOL helps with other types of reading comprehension abilities.

Last but not least, we also tested the models pre-trained with IDOL on MNLI (Williams et al.,

[3] In the following sections, we refer these baseline models to BERT, RoBERTa, ALBERT and DeBERTa respectively for simplicity.

2018) and STS-B (Cer et al., 2017), two tasks of GLUE (Wang et al., 2018), to make sure that the general language understanding abilities are retained to a great extent during the process of logical enhancement. The evaluation metrics on STS-B are the Pearson correlation coefficient (Pear.) and Spearman's rank correlation coefficient (Spear.) on the development set. And we use the accuracy of MNLI-m and MNLI-mm development sets for evaluation on MNLI.

ReClor The problems in this dataset are collected from two American standardized tests - LSAT and GMAT, which guarantee the difficulty of answering the questions. Moreover, ReClor covers 17 classes of logical reasoning including main idea inference, reasoning flaws detection, sufficient but unnecessary conditions, and so forth. Each problem consists of a passage, a question, and four answer candidates, like the one shown in the green section of Figure 1. There are 4638, 500, and 1000 data points in the training set, development set, and test set respectively. The accuracy is used to evaluate the system's performance.

LogiQA The main difference compared with ReClor is that the problems in LogiQA are generated based on the National Civil Servants Examination of China. Besides, it incorporates 5 main reasoning types such as categorical reasoning and disjunctive reasoning. And 7376, 651, and 651 samples are gathered for the training set, development set, and test set individually.

5.3 Implementation Detail

5.3.1 IDOL

During the process of pre-training with IDOL, we implemented the experiments on 8 Nvidia A100 GPUs. Since IDOL was applied on multiple different pre-trained models, we provide a range for some main hyperparameters. The whole training process consists of 10k~20k steps while the warm-up rate keeps 0.1. The learning rate is warmed up to a peak value between 5e-6~3e-5 for different models, and then linearly decayed. As for batch size, we found that 1024 or 2048 is more appropriate for most models. Additionally, we use AdamW (Loshchilov and Hutter, 2017) as our optimizer with a weight decay of around 1e-3. For the software packages we used in detail, please see Appendix.

With respect to the hyperparameters for fine-tuning models on downstream tasks, we follow the configurations provided in the original paper of either the corresponding model or the dataset.

Models	ReClor		LogiQA	
	Dev	Test	Dev	Test
BERT	53.8	49.8	35.3♠	33.0♠
IDOL	**56.8**	**53.3**	**36.9**	**34.3**
RoBERTa	62.6	55.6	37.0♠	36.6♠
DAGN	65.2	58.2	35.5	38.7
AdaLoGN	65.2	60.2	39.9	40.7
LReasoner	66.2	62.4	38.1	40.6
MERIt	67.8	60.7	42.4	41.5
Logiformer	68.4	63.5	42.2	**42.6**
IDOL	**70.2**	**63.9**	**42.5**	41.8
ALBERT	70.4	67.3	41.2♠	41.3♠
LReasoner	73.2	70.7	41.6	41.2
MERIT	73.2	**71.1**	43.9	**45.3**
IDOL	**74.6**	70.9	**44.7**	43.8

Table 2: Results on logical reasoning MRC benchmarks - ReClor and LogiQA. In each block, the previous methods listed for comparison and IDOL take the pre-trained model in the first line as their backbone model. ♠: reproduced by ourselves.

5.3.2 LLM

For the purpose of comparing IDOL with LLMs, we randomly sampled 30 pieces of data in the development sets of ReClor and LogiQA separately (named Dev-30). As for models, we choose GPT-3.5[4], ChatGPT[5] and GLM-130B (Zeng et al., 2022) for this pilot test.

To better evaluate the performance of LLMs, we tested them in the following three settings: zero-shot prompting, few-shot prompting, and chain-of-thought prompting. For zero-shot prompting, we designed the following template to wrap up the MRC problem.

> *The passage is* `[PASSAGE]`. *The question is* `[QUESTION]`. *Here are 4 choices for it and they are* `[CHOICES]`. *Which one should I choose? Thanks.*

As for few-shot prompting, we insert 3 examples in the same template but with correct answers ahead of the target question. When testing with chain-of-thought prompting, the template is similar to the one presented above. But there is only one example ahead and sentences describing the

[4]The exact version is text-davinci-003.

[5]Tested on February 13th, 2023.

Models	ReClor		
	Test	Test-E	Test-H
DeBERTa♡	75.3	84.0	68.4
LReasoner♣	76.1	87.1	67.5
Knowledge Model♣	79.2	**91.8**	69.3
MERIt♣	79.3	85.2	74.6
AMR-LE♣	80.0	87.7	73.9
IDOL	**80.6**	87.7	**75.0**

Table 3: Results of IDOL with DeBERTa and other publicly available data. ♣: top results from the official leaderboard of ReClor (as of January 19, 2023). ♡: the performance of the original DeBERTa from Jiao et al. (2022) for reference (the majority of the top submissions and IDOLs in this table take DeBERTa as the backbone model).

process of the way how humans reason to solve the problem are provided before giving the right answer to the example. For more details about the templates and the test example, please refer to Table 6 and Figure 4.

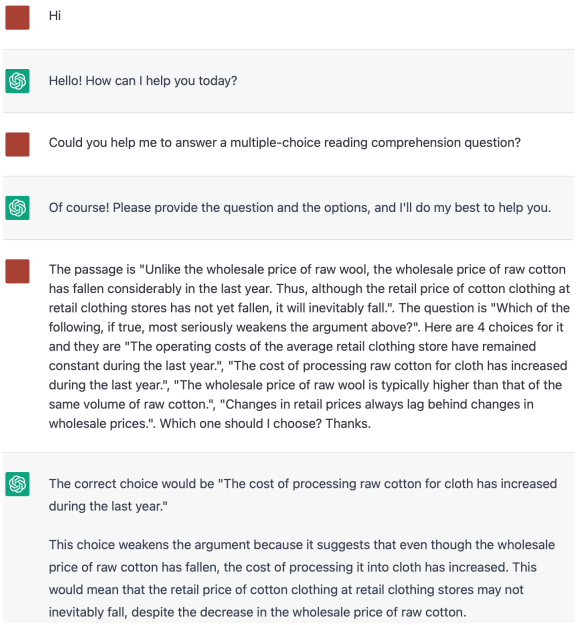

Figure 4: An example of ChatGPT answering an MRC question.

5.4 Main Results

5.4.1 Logical Reasoning MRC

Fine-tuning To evaluate the model performance on logical reasoning MRC, we experimented with the baseline models mentioned above on ReClor and LogiQA, the two most representative benchmarks in this field. The majority of previous researchers focus on applying their method to RoBERTa, IDOL meets the most competitors in this setting as shown in Table 2. In spite of this, IDOL surpassed all the existing strong systems by an obvious margin in nearly every evaluation metric except the accuracy on the LogiQA test set. Apparently, from the results on BERT and ALBERT in Table 2 and results on DeBERTa in Table 3, we can see that IDOL has significant advantages over other opponents as well. In summary, IDOL is highly effective in logical reasoning MRC with state-of-the-art performance and this benefit can be generalized to different pre-trained models even to the recent large-scale and strong ones.

Prompting Although the scale of Dev-30 for the pilot test on LLM is small, the results displayed in Table 5 inspired us to some extent. Generally, IDOL is still competitive in the era of LLM. On ReClor, it achieved an accuracy of 80% while the best result from LLMs is 70% (ChatGPT with Chain-of-Thought prompting). Even though GLM-130B realizes an accuracy of 50% on LogiQA in Zero-Shot setting surprisingly (slightly higher than 43.3% by IDOL), IDOL has an obvious advantage compared with other settings and other LLMs. Additionally, there is an interesting phenomenon that chain-of-thought prompting brings negative effects on LLMs except for ChatGPT on ReClor, which is not consistent with the findings in Wei et al. (2022).

5.4.2 Other MRC Datasets

For testing whether IDOL could also benefit on types of MRC tasks or maintain the original abilities, we conducted a series of experiments based on RoBERTa as the backbone model. The results are displayed in the middle part of Table 4 where we compare the original model, the model further pre-trained with only MLM on LGP and the model further pre-trained with IDOL. We evaluate the models in each setting with 4 different seeds and report the average value. It is apparent that IDOL performs better on both RACE and SQuAD 2.0 in each evaluation metric (although the effects are not as big as those on ReClor or LogiQA), which implies that IDOL indeed helps on general MRC tasks while achieving significant improvement in logical reasoning ability.

5.4.3 General Understanding Ability

Following the experiment configuration in section 5.4.2, we planned to find out what kind of effect would IDOL have on other types of natural lan-

Models	ReClor		LogiQA		RACE		SQuAD 2.0		STS-B		MNLI	
	Dev	Test	Dev	Test	Dev	Test	F1	EM	Pear.	Spear.	m	mm
RoBERTa	62.7	55.2	36.2	37.1	85.2	84.4	89.0	86.1	**92.6**	**92.5**	89.5	89.3
+MLM	65.0	58.4	37.9	36.6	85.4	84.5	89.0	86.1	92.2	92.1	89.5	89.5
+LCP (IDOL)	**66.8**	**60.6**	**39.4**	**38.8**	**85.6**	**84.8**	**89.2**	**86.2**	92.3	92.2	**89.7**	**89.5**

Table 4: Results of RoBERTa with different pre-training tasks on logical reasoning MRC, other types of MRC and other types of NLU tasks.

Models	Setting			
	ZS	FS	CoT	FT
ReClor				
GPT-3.5	56.7	50.0	46.7	-
ChatGPT	63.3	63.3	70.0	-
GLM-130B	46.7	40.0	23.3	-
IDOL	-	-	-	**80.0**
LogiQA				
GPT-3.5	30.0	10.0	13.3	-
ChatGPT	33.3	36.7	23.3	-
GLM-130B	**50.0**	36.7	26.6	-
IDOL	-	-	-	43.3

Table 5: Results on ReClor and LogiQA from LLMs and IDOL. ZS: Zero-Shot prompting. FS: Few-Shot prompting. CoT: Chain-of-Thought prompting. FT: Fine-Tuning.

guage understanding tasks which help to reflect the general understanding ability of pre-trained language models. We evaluate the models in each setting with 4 different seeds and report the average value. From the results presented in the right part of Table 4, we can easily find that although IDOL falls behind on MNLI and exceeds the other two competitors on STS-B, the differences in all the evaluation metrics are quite small. Therefore, we could conclude that IDOL retains the general language understanding ability from the original pre-trained model successfully during the process of becoming stronger in logical reasoning.

6 Ablation Study

In this section, we conducted a series of ablation experiments about the multiple logical indicators we used in both fine-tuning and pre-training phases. We evaluate the models based on RoBERTa with 4 different seeds and report the average value.

6.1 Indicators in Fine-tuning

As introduced in section 3.2, we defined 5 classes of logical indicators that reflect various logical relations among text logical units and we make use of all of them in IDOL. To figure out whether the 5 types are of equal importance in logical reasoning MRC, we conducted a set of controlled experiments where certain types of indicators are removed from the ReClor train set as the fine-tuning train dataset in each setting.

From the results displayed in Table 7, it is obvious from the last column that logical indicators indeed play an important role in logical reasoning-related text understanding since the loss of all indicators decreases accuracy by 4 to 7 points. In detail, we can conclude that the negative and adversative indicators influence the most by comparing the gaps between pre-training on the original LGP and the dataset without individual types of indicators.

6.2 Indicators in Pre-training

Now that logical indicators have been proven to be effective in fine-tuning stage, we believe they also help with the pre-training stage. Therefore, we arranged a series of experiments on gradually incorporating more logical indicators from not leveraging any indicators (MLM), only making use of PMI and CLI (LCP-2), adding LUI to LCP-2 (LCP-3), to taking advantage of all 6 types of logical indicators (LCP).

From the lines displayed in Figure 5, it is clear that models perform better while leveraging a greater variety of logical indicators since the red line (IDOL) is positioned significantly higher than green and yellow lines representing pre-training tasks that utilize fewer types of logical indicators. According to the results in Table 7, PMI and CLI brought the least difference in the model performance on ReClor. The LCP-2 and LCP-3 mainly rely on the two types, and introducing a new special

Setting	Template
Zero-Shot	*The passage is* `[PASSAGE]`. *The question is* `[QUESTION]`. *Here are 4 choices for it and they are* `[CHOICES]`. *Which one should I choose? Thanks.*
Few-Shot	`[Example A]` `[Example B]` `[Example C]` *The passage is* `[PASSAGE]`. *The question is* `[QUESTION]`. *Here are 4 choices for it and they are* `[CHOICES]`. *Which one should I choose? Thanks.*
Chain-of-Thought	*The passage is* `[PASSAGE]`. *The question is* `[QUESTION]`. *Here are 4 choices for it and they are* `[CHOICES]`. *You can analyze like this,* `[Thought Process]`. *So the answer is* `[Answer]`. *The passage is* `[PASSAGE]`. *The question is* `[QUESTION]`. *Here are 4 choices for it and they are* `[CHOICES]`. *Which one should I choose? Thanks.*

Table 6: Templates and examples for LLM prompting in different settings.

Models	ReClor Train Set					
	—	PMI&CLI	NTI	ATI	CNI	ALL
RoBERTa	62.7	64.0	59.7	61.7	63.7	59.1
+ MLM	65.0	64.9	61.8	61.5	64.5	59.9
+ LCP	66.8	63.8	62.7	63.4	64.2	60.5

Table 7: Results of fine-tuning with datasets obtained by removing certain types of logical indicators in the original ReClor train set and testing on the development set. The first row under "ReClor Train set" in each column indicates what indicators are removed from LGP. "—": the original LGP. "PMI&CLI": both premise and conclusion indicators are removed. "ALL": no logical indicators left.

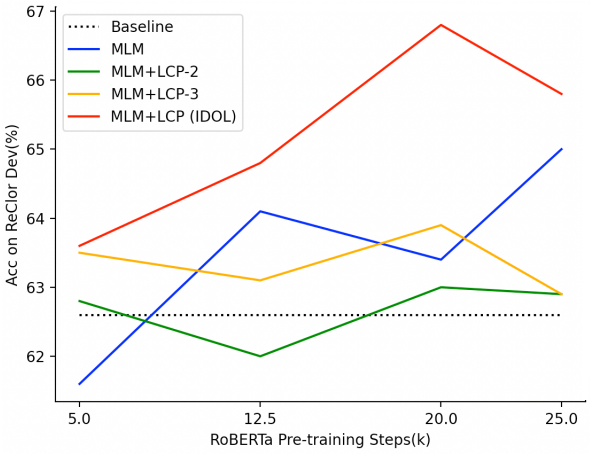

Figure 5: The results on ReClor development set of models with different tasks on RoBERTa during the pre-training. LCP-2: LCP only with PMI and CLI. LCP-3: LCP only with PMI, CLI, and LUI. Baseline: fine-tuning with the original RoBERTa.

token [LGMASK] inevitably brings noise during model training and further widens the gap between pre-training and down-stream tasks, so that they perform even not better than the original MLM. Additionally, in the aspect of overall trends, the model pre-trained with IDOL is becoming stronger gradually during the process of pre-training, which certifies the effectiveness of our designed task targeted at logical indicators.

7 Conclusion and Future Work

In this paper, we proposed an easy-to-understand further pre-training method IDOL which fully exploits the logical information provided by 6 types of logical indicators and is proven effective on different pre-trained language models while keeping them competitive on many other kinds of down-stream tasks. Particularly, IDOL achieves state-of-the-art performance on logical reasoning machine reading comprehension tasks.

With respect to future work, we plan to leverage the sentence-level or passage-level logical features in the meantime and integrate it with IDOL to generate a stronger multi-task further pre-training method for improving the logical reasoning ability of pre-trained language models. Moreover, we decide to redesign the IDOL task and find out whether logical indicators also play an important role in those generative pre-trained models as well. Furthermore, we will explore the way of combining IDOL with prompting to find a better method to elicit the reasoning abilities of LLMs.

8 Limitations

First of all, IDOL relies on a customized dataset that is filtered out from Wikipedia pages with the help of many pre-defined logical indicators. Inevitably, this will introduce a certain amount of artificial bias. If an automatic method for logical indicator extraction based on something like hidden representations from neural network models is put forward, it would be beneficial to narrow the gap between the dataset preparation and logical pre-training.

In addition, in the field of pre-training task design, there have been a lot of different but effective approaches proposed. For example, in Cui et al.

(2022), the authors presented a pre-training task named PERT which requires the models to recover the original token sequences under the background of that different token permutation within a certain range would not affect Chinese text understanding. This method only depends on the original texts, but IDOL introduces one more special token, which widens the gap between pre-training and fine-tuning to some extent.

References

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Ting Liu. 2022. Pert: Pre-training bert with permuted language model.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Chadi Helwe, Chloé Clavel, and Fabian M. Suchanek. 2021. Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction*.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. DAGN: Discourse-aware graph network for logical reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5848–5855, Online. Association for Computational Linguistics.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3496–3509, Dublin, Ireland. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Xiao Li, Gong Cheng, Ziheng Chen, Yawei Sun, and Yuzhong Qu. 2022. AdaLoGN: Adaptive logic graph network for reasoning-based machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7147–7161, Dublin, Ireland. Association for Computational Linguistics.

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 *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization. Main track.

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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xinyu Pi, Wanjun Zhong, Yan Gao, Nan Duan, and Jian-Guang Lou. 2022. Logigan: Learning logical reasoning via adversarial pre-training. *ArXiv*, abs/2205.08794.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2022a. Logic-driven context extension and data augmentation for logical reasoning of text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1619–1629, Dublin, Ireland. Association for Computational Linguistics.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2022b. Logic-driven context extension and data augmentation for logical reasoning of text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1619–1629, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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

Fangzhi Xu, Jun Liu, Qika Lin, Yudai Pan, and Lingling Zhang. 2022. Logiformer: A two-branch graph transformer network for interpretable logical reasoning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1055–1065, New York, NY, USA. Association for Computing Machinery.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
We just use Grammarly to do spell checks.

B ☑ Did you use or create scientific artifacts?

4.1

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will put forward the terms for using the artifact we created when we publish it, only for research purposes.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We will put forward the terms for using the artifact we created when we publish it, only for research purposes.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
As far as we know, something like names is safe in the Wikipedia corpus and there is nearly no offensive content in it, so we didn't plan to filter out those texts like names or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5.4 and 6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.