LogiQA 2.0 — An Improved Dataset for Logical Reasoning in Natural Language Understanding

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, Yue Zhang†

Abstract—NLP research on logical reasoning regains momentum with the recent releases of a handful of datasets, notably LogiQA and Rectr. Logical reasoning is exploited in many probing tasks over large Pre-trained Language Models (PLMs) and downstream tasks like question-answering and dialogue systems. In this paper, we release LogiQA 2.0. The dataset is an amendment and re-annotation of LogiQA in 2020, a large-scale logical reasoning reading comprehension dataset adapted from the Chinese Civil Service Examination. We increase the data size, refine the texts with manual translation by professionals, and improve the quality by removing items with distinctive cultural features like Chinese idioms. Furthermore, we conduct a fine-grained annotation on the dataset and turn it into a two-way natural language inference (NLI) task, resulting in 35k premise-hypothesis pairs with gold labels, making it the first large-scale NLI dataset for complex logical reasoning. Compared to Question Answering, Natural Language Inference excels in generalizability and helps downstream tasks better. We establish a baseline for logical reasoning in NLI and incite further research.

Index Terms—Reading Comprehension, Logical Reasoning, Natural Language Inference, Textual Inference

I. INTRODUCTION

The capability of logical reasoning is a crucial part of natural language understanding (NLU) [1] [2] [3]. Investigation of linguistic reasoning dates back to the 1950s, at the dawn of computer science and artificial intelligence [4] [5] [6] [7] [8]. However, with limited computing power and primitive NLU technologies, formal logical reasoning gradually dominated the research field in the 1970s and became a key area of AI research over a long period [9] [10].

Recently, with the advance of deep learning technology, NLU has witnessed significant improvements [13] [14], with competitive results being reported over typical tasks, including natural language inference (NLI) [15] [16] and machine reading comprehension (MRC) [17] [18]. Figure 1 illustrates the two NLU tasks. In Figure 1(a), an NLI model takes the premise and hypothesis as input and predicts whether the premise entails the hypothesis. In Figure 1(b), an MRC model takes a passage and question pair as input to predict the correct answer. There is a fundamental connection between machine reading comprehension and natural language inference [15], both tasks rely heavily on reasoning skills, and both are general because many NLP tasks can be cast into MRC [19] [20] or NLI [21]. For both NLI and MRC tasks, the current state-of-the-art approaches make use of a sizeable pre-trained language model such as BERT [13], and RoBERTa [22], fine-tuned using the benchmark-specific training data. Benefiting from large-scale pre-training, such models have achieved performances close to or surpass the human level on popular benchmarks [23] [24].

The recent advance in NLU leads to the natural question of whether it is time to revisit traditional linguistic reasoning tasks. Relevant to this question, some work has shown evidence that the current deep learning technologies have the potential to conduct logical reasoning [25]. From the application perspective, harnessing logical reasoning benefits downstream tasks and NLP applications, such as dialogue systems [26], information extraction [27], and question answering [28] [29]. However, for both NLI and MRC, most existing datasets are designed to evaluate the capabilities of basic linguistic understanding, as Figure 1 shows. Relatively few benchmarks are available for systematically measuring the performance of...
There are four cups on the table, each with a sentence written on it. The first cup: “Beer is in all cups.” The second cup: “Cola in this cup.” The third cup: “No coffee in this cup.” Fourth cup: “Some cups have no beer.” Only one of the four options is explicitly described in the context. In contrast, for traditional MRC tasks, such as SQuAD [12], and HotpotQA [35], only explicit evidence integration is necessary. LogiQA contains 8,678 paragraph-question pairs, each with four candidate answers. Using this dataset, we evaluated the capacity of pre-trained language models, in particular, BERT [13] and RoBERTa [22], for logical MRC. Results show a significant gap between model performance (around 35%) and human level (around 86%), revealing the shortcomings of pre-trained LMs despite their success on traditional datasets. Our dataset facilitated subsequent research on the critical examination of existing datasets [36] [37], investigating various reasoning skills [38] [39], and designing new neural structures for language models [40].

The original LogiQA dataset, however, has three noteworthy limitations. First, LogiQA 1.0 categorizes 651 test samples into five reasoning types, namely categorical reasoning, sufficient conditional reasoning, necessary conditional reasoning, disjunctive reasoning, and conjunctive reasoning. However, it does not categorize the whole dataset, limiting the use of the dataset for investigating a sub-category of challenges in isolation or for fine-grained evaluation of reasoning capabilities. Second, as the original LogiQA dataset has been highly challenging to neural models [41], with the best-reported results being 39.32% accuracy [40], leaving a steep curve for research on more effective models. To this end, a binary NLI classification task can potentially reduce the ambiguity in the output from 4 choices to 2, which offers a different perspective to logical NLU. However, NLI is not included in LogiQA 1.0. Third, the quality of the English dataset needs improving because it has been translated using machine translation and human post-editing. One example of the test input is shown in Figure 4(a). As can be seen, the ambiguity of the original translation of option B undermines the truth value. As a result, the quality of translation can affect the effectiveness 1 of the dataset as an effective benchmark for evaluating logical reasoning capabilities.

We present LogiQA 2.0, an improved version of the LogiQA dataset. Figure 3 shows examples of two tasks in the LogiQA 2.0 dataset. The LogiQA 2.0 dataset has four salient changes.

The Chinese dataset is not affected as it is entirely human-curated.
First, we enlarged the dataset from 8678 instances to 15708 instances by collecting more logical test materials. Second, as shown in Figure 4(b), the quality of the English data is largely improved with fully professional human translation. Third, an NLI dataset is further added by making adaptations to the original MRC instances. Fourth, we give a fine-grained analysis of different logical reasoning types according to the division of logical reasoning types.

We reran typical baseline state-of-the-art NLP models on LogiQA 2.0, finding that the MRC task achieves better performance. For LogiQA 2.0, BERT-base achieves 48.12% accuracy, compared to 33.83% on LogiQA 1.0. We further quantify the influence of the English translation quality and the size of data, both of which have theoretical and practical significance. For the logical NLI task, which is not included in LogiQA 1.0, models achieve an accuracy of 57.36%. with BERT-base achieving 54.97% accuracy. In addition to BERT-base and RoBERTa, which we experimented with in LogiQA 1.0, we additionally evaluate BART [14], GPT-2 [42] and GPT-3 [43] for both LogiQA 1.0 and LogiQA 2.0, which represent pre-trained models of different architectures. The best-performing GPT-3 gives results of 54.93% and 68.65% on LogiQA 2.0 MRC and NLI, respectively. Though much higher compared with BERT-base, the results lag much behind the 90% human level, which shows that logical reasoning is a challenging issue for both tasks.

This article is a significant extension to our LogiQA 1.0 conference paper [44], with rewritten introduction and related work sections, a new dataset release, updated results, and extended experimental discussion details. All the resources are released at https://github.com/csitfun/LogiQA2.0.

### II. DATASET

This section describes the LogiQA 2.0 dataset, a superset of LogiQA 1.0 with quantity and quality improvements. For brevity, we do not re-introduce LogiQA 1.0 but give the major differences between the two versions in Section II-E.

#### A. MRC Data Collection and Translation

Our data source is public materials from the Chinese Civil Service Exam, which is held annually in provinces of China 2. Experts design formal logical reasoning questions to test candidates’ critical thinking and problem-solving skills. The exams take the form of multi-choice question answering, and testees need to select the correct options to answer the question. The testing materials are released online after each year’s examination. LogiQA 2.0 increases the data size by including more logical reasoning tests.

We collect our data from the web 3, obtaining 16490 text-question-options triples. Each item contains one paragraph of text, a question, and four options. The correct answer or answers are labeled for each item. Discrimination is provided during collecting to rule out instances that are not self-contained by the texts, i.e., instances are neglected if they contain charts or tables; we did not include questions that need to fill in the blanks. After the data cleaning process, the resulting dataset contains 15937 paragraph-questions pairs. We release the data as a Chinese MRC corpus for LogiQA 2.0.


URL links of the data sources can be found in our dataset repository. https://github.com/csitfun/LogiQA2.0-Chinese

---

**Fig. 5:** Examples of each type of logical reasoning in LogiQA 2.0. (✓) indicates the correct answer.
For LogiQA 2.0, we then set up a translation project to translate the dataset into English, along with further data cleaning procedures. Professional translators are hired to translate the dataset from Chinese to English. Through this process, known issues of all the instances in LogiQA 1.0 are manually fixed. We remove all the instances that center on ancient poetry, Chinese philology, etc. Examples are given in Appendix A). In total, 667 translator hours (about 20 problems per hour) are spent with an average pay rate of 67 RMB ($10 USD) per hour. Figure 3(b) in the introduction shows one instance of the LogiQA 2.0 MRC task.

### B. The NLI Section of LogiQA 2.0

The NLI task was first introduced in the PASCAL Recognizing Textual Entailment Challenge [15] in the 2000s. The task is to decide the entailment relationship between two text fragments and has a natural connection with logical reasoning. Researchers used expert-designed datasets to explore different aspects of textual entailment [16], [45]. Crowdsourced methods were also introduced to construct large-scale NLI datasets [46], [47] in the 2010s. A typical example of the NLI datasets is shown in Figure 1(a). In general, the process of generating high-quality NLI datasets is costly. Due to the scarcity of NLI datasets, work has been done to obtain golden labeled NLI data from other NLU tasks, such as question-answering [48], [49], and summarization [50].

Following the work of Demszky et al. [49], we use rule-based method to extract the premise and hypotheses from the concatenation of text, question, and options of each MRC instance in LogiQA 2.0. As an NLI dataset for logical reasoning, the NLI section of LogiQA 2.0 complies with the formulation of classical NLI tasks. It decides the entailment relationship between two text fragments (one is called the premise, and the other is the hypothesis) with a two-way classification scheme.

As shown in Figure 3(a), because the premise of each NLI instance is at the multi-sentence level, we follow the concept of syllogism [51] [52] and further add fine-grained labels to the premise, differentiating its major components. Such labels can optionally be used in training NLI models, which are illustrated in the coloring of Figure 3(a). One similar NLI dataset is the FraCaS Textual Inference Problem Set [16], derived from the FraCaS project [53]. There are 346 problems, each containing one or more premises and a hypothesis. Below is an illustrated example:

P1 A Swede won a Nobel prize.

P2 Every Swede is a Scandinavian.

H A Scandinavian won a Nobel prize.

In this example, categorical reasoning is the key reasoning skill to decide the entailment relation. P2 is the major premise [54], a statement of a general or universal nature; P1 is the minor premise, a statement regarding a particular case, related to the subject of the major premise [55]. Compared to the FraCaS dataset, our NLI section of the LogiQA 2.0 is larger in data size, and the context of the premise is longer.

Such additional fine-grained labels are designed to suit the setting of logical reasoning in the NLI setting, and can potentially facilitate research in syllogism (conditional, categorical, disjunctive) reasoning. Our labeling is done using hand-crafted rules. As shown in Figure 3(a), each item is annotated with one of the two labels, Entailed and Not Entailed. One premise may have multiple hypotheses.

The resulting dataset consists of 39,414 premise-hypothesis pairs, covering a broad spectrum of logical reasoning types and a large genre of topics. Figure 3(a) in the introduction shows one instance of the LogiQA 2.0 NLI task. To the best of our knowledge, the NLI version of LogiQA 2.0 is the first large-scale expert-designed logical reasoning dataset for NLI beyond the sentence level featuring optional fine-grained premise labeling.

### C. Reasoning Types

For both the MRC section and the NLI section of our dataset, we categorize the instances according to the five types of formal logical reasoning defined by Hurley [56], which are categorical reasoning, sufficient conditional reasoning, necessary conditional reasoning, disjunctive reasoning, and conjunctive reasoning. Human workers manually assign all the reasoning type labels based on the reasoning type definitions. The reasoning type of each question is assigned by a total of 5 workers, and each of them corresponds to one reasoning type. We give the description of reasoning types to the workers. The reasoning type of each question is a collection of 5 workers’ decisions. The representative examples of different reasoning types and their proportion are illustrated in Figure 5. Note that the sum of total values is bigger than 100% for the reason that one instance may involve multiple reasoning types. These types of reasoning belong to deductive reasoning, for which a definite conclusion can be derived given a set of premises. As a result, such reasoning can be most suitable for evaluating performances quantitatively. Formally, the five types of reasoning can be described as follows:

5 https://github.com/ni-li-for-qa/conversion

6 An illustration is shown in Appendix B
- **Categorical reasoning:** The goal is to reason whether a specific concept belongs to a particular category. This type of reasoning is commonly associated with quantifiers such as “all/everyone/any”, “no”, and “some”, etc.

- **Sufficient conditional reasoning:** This type of hypothetical reasoning is based on conditional statements of the form “If \( P \) then \( Q \)”, in which \( P \) is the antecedent and \( Q \) is the consequent.

- **Necessary conditional reasoning:** This type of hypothetical reasoning is based on conditional statements of the form “\( P \) only if \( Q \)”, “\( Q \) whenever \( P \)”, etc., where \( Q \) is a necessary condition for \( P \).

- **Disjunctive reasoning:** In this type, the premise is disjunctive, in the form “either . . . or . . .”, where the conclusion holds as long as one premise holds.

- **Conjunctive reasoning:** In this type of reasoning, the premise is conjunctive, in the form “both . . . and . . .”, where the conclusion holds only if all the premises hold.

### D. Data Statistics

The detailed statistics of the LogiQA 2.0 MRC task are summarized on the left of Table I. We have 15708 paragraph-question pairs in the MRC section of our dataset. The average paragraph length is relatively small compared with existing reading comprehension datasets since logical reasoning problems do not rely heavily on complex context. We randomly split the dataset, using 12567, 1569, and 1572 instances for training, development and testing, respectively.

The statistics of the LogiQA 2.0 NLI task are summarized on the right of Table I. It is converted from the MRC section and has 39414 premise-hypothesis pairs. Compared to traditional one-sentence NLI datasets, the average premise length is 71.65, much longer due to the multi-sentence context. We randomly split the dataset, using 31531 instances for training, 3941 instances for development, and 3942 instances for testing.

### E. LogiQA 2.0 versus LogiQA 1.0

The main differences between LogiQA 2.0 and LogiQA 1.0 are 4-fold. The first is the data size. LogiQA 1.0 contains 8678 question instances, and LogiQA 2.0 extended it to 15,708 question instances. The second is the task framework. LogiQA 1.0 is a dataset solely for MRC in the form of four-option question answering; LogiQA 2.0 added a two-way NLI task with 39,414 premise-hypothesis pairs, making it a dataset containing multiple tasks. The third is data quality. We enhanced the data quality with retranslation and amendments; elusive instances that may hinder the understanding of broader audiences are removed from the dataset. The fourth is more elusive instances that may hinder the understanding of broader audiences.

### III. NLI Models

We use the state-of-the-art pre-trained language models to establish baselines for LogiQA 2.0 and empirically understand such models’ effectiveness on logical NLI. In addition, human performance is evaluated as a reference metric.

#### A. Pre-trained Language Models

We take the following pre-trained models:

- **BERT** [13] is a Transformer-based [62] language model that uses a masked language modeling objective.

- **RoBERTa** [22] extends BERT using a more dynamic sentence masking method. It uses a similar architecture as BERT but has a different pre-training scheme.

- **XLNet** [60] is a generalized autoregressive pre-training method that uses a permutation language modelling objective to combine the advantages of autoregressive and autoencoding methods.

- **BART** [14] is a sequence-to-sequence model with a denoising bidirectional autoencoder and a left-to-right decoder. The BART model works both for text generation tasks and comprehension tasks.

- **GPT-2** [61] is the second generation of the Generative Pre-Trained (GPT) [63] language model that uses the decoder part of the transformers architecture. GPT-2 is trained on a very large corpus of English texts with a “predict the next word” unsupervised fashion.

- **GPT-3** [43] is the third generation of GPT; it is an autoregressive language model that can generate human-like texts. The architecture of GPT-3 is based on the transformer network with a generative training objective on Internet-scale text data.

We apply fine-tuning for BERT, RoBERTa, XLNet, BART, and GPT-2. Following Devlin et al., given a premise \( p \) and a hypothesis \( h \), we concatenate the premise-hypothesis pair as

```

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev Accuracy</th>
<th>F1-score</th>
<th>Test Accuracy</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>54.97</td>
<td>57.39</td>
<td>54.87</td>
<td>59.61</td>
</tr>
<tr>
<td>RoBERTa-base</td>
<td>55.41</td>
<td>56.13</td>
<td>55.43</td>
<td>57.96</td>
</tr>
<tr>
<td>XLNet-base</td>
<td>55.42</td>
<td>58.78</td>
<td>55.89</td>
<td>60.12</td>
</tr>
<tr>
<td>BART-base</td>
<td>56.83</td>
<td>57.41</td>
<td>57.08</td>
<td>56.89</td>
</tr>
<tr>
<td>GPT-2-base</td>
<td>55.01</td>
<td>57.84</td>
<td>54.75</td>
<td>57.32</td>
</tr>
<tr>
<td>GPT-3</td>
<td>66.49</td>
<td>69.55</td>
<td>67.78</td>
<td>68.65</td>
</tr>
</tbody>
</table>

```

### TABLE III: Results on LogiQA 2.0 NLI (accuracy%).

We apply fine-tuning for BERT, RoBERTa, XLNet, BART, and GPT-2. Following Devlin et al., given a premise \( p \) and a hypothesis \( h \), we concatenate the premise-hypothesis pair as

```

<table>
<thead>
<tr>
<th>Human Performance</th>
<th>86.63</th>
<th>84.06</th>
<th>89.36</th>
<th>91.40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceiling Performance</td>
<td>98.00</td>
<td>99.31</td>
<td>96.67</td>
<td>97.97</td>
</tr>
</tbody>
</table>

```
a new sequence $[CLS] + p + [SEP] + h + [SEP]$, where $[CLS]$ and $[SEP]$ are special symbols for classification token and separator token, respectively. After encoding, the last layer’s hidden representation from the $[CLS]$ token is fed in an MLP + softmax for classification. For the encoder-decoder models BART, we take $[CLS]$ from the decoder module. For the decoder model GPT-2, the last token $[SEP]$ is used for classification instead of the first $[CLS]$ token. In addition to the standard NLI setting, we conduct separate experiments to examine the role of additional fine-grained premise labels. We add the leading symbol $[Major premise]$ before the major premise texts, and $[Minor premise]$ before the minor premise texts, resulting in the sequence $[CLS] + [Major premise] + p_{major} + [Minor premise] + p_{minor} + [SEP] + h + [SEP]$ as the input.

For the GPT-3 model, we use in-context learning to explore GPT-3’s few-shot learning ability by using a few examples. Specifically, we use 5 examples, each of which represents a typical reasoning skill as discussed in Section II-C. For the NLI task, each example is organized in a uniform format:

**Given the fact:** {premise}

**Does it follow that:** {hypothesis}

**Yes or no?** {label}

**B. Human Performance**

To measure human performance on the LogiQA 2.0 NLI dataset, we randomly select 300 context-hypothesis pairs from the test set. Four testees were recruited, who are master students in computer science, and two had the experience of preparing for the Chinese Civil Service Exam. We report human performance by the mean score. The ceiling performance is obtained by considering the proportion of questions with at least one correct answer.

<table>
<thead>
<tr>
<th>Category</th>
<th>Model</th>
<th>LogiQA 1.0</th>
<th>LogiQA 2.0</th>
<th>Chinese LogiQA 1.0</th>
<th>Chinese LogiQA 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
</tr>
<tr>
<td>Rule-based</td>
<td>Random(theroretical)</td>
<td>25.00</td>
<td>25.00</td>
<td>25.00</td>
<td>25.00</td>
</tr>
<tr>
<td>Pre-trained</td>
<td>BERT [13]</td>
<td>33.83</td>
<td>32.08</td>
<td>48.12</td>
<td>47.90</td>
</tr>
<tr>
<td></td>
<td>RoBERTa [22]</td>
<td>35.85</td>
<td>35.31</td>
<td>49.88</td>
<td>48.76</td>
</tr>
<tr>
<td></td>
<td>XLNet [60]</td>
<td>35.02</td>
<td>33.17</td>
<td>50.86</td>
<td>47.84</td>
</tr>
<tr>
<td></td>
<td>BART [44]</td>
<td>35.49</td>
<td>35.61</td>
<td>48.80</td>
<td>46.31</td>
</tr>
<tr>
<td></td>
<td>GPT-2 [61]</td>
<td>33.93</td>
<td>33.48</td>
<td>47.05</td>
<td>47.32</td>
</tr>
<tr>
<td></td>
<td>GPT-3 [43]</td>
<td>51.53</td>
<td>49.17</td>
<td>56.21</td>
<td>54.93</td>
</tr>
<tr>
<td>Human</td>
<td>Human Performance</td>
<td>-</td>
<td>86.00</td>
<td>-</td>
<td>84.00</td>
</tr>
<tr>
<td></td>
<td>Ceiling Performance</td>
<td>-</td>
<td>95.00</td>
<td>-</td>
<td>98.00</td>
</tr>
</tbody>
</table>

**TABLE IV:** Main results on LogiQA 2.0 MRC task (accuracy%).

**B. Pre-trained Language Methods**

We take the same pre-trained models discussed in Section III-A for our experiments.

For BERT, RoBERTa, XLNet, BART, and GPT-2, we concatenate the paragraph, question, and each candidate answer as one sentence. Then use a pre-trained contextualized embedding model to encode the sentence for calculating its score. In the multi-choice scenario, as four candidate answers in our dataset, four concatenated sentences are constructed by pairing each candidate answer with the paragraph and question. The one with the highest model score is chosen as the answer. In particular, we treat the paragraph as sentence A and the concatenation of the question and each candidate as sentence B. For GPT-3, because of the few-shot learning scenario, we use the 5 typical examples discussed in Section II-C as in-context, each example is organized in a uniform format:

**Given the fact:** {text}

**Question:** {option1} / {option2} / {option3} / {option4}

**Answer:** {answer}

**C. Human Performance**

To measure human performance on the LogiQA 2.0 MRC dataset, we randomly select 100 MRC instances from the test set. The number of instances is proportional to that of the NLI section. We use the same group of testees as the NLI section for measuring human performance. The human performance is reported by average scores, and the human ceiling performance is obtained by considering the proportion of questions with at least one correct answer.

**V. Experiments**

We report the main results of the aforementioned models in Section III and IV on the LogiQA 2.0 NLI and MRC sections, respectively, to make a comparison between the performances.
of different types of models, and between the state-of-the-art models and human level. In addition, for the MRC section, we also compare results on LogiQA 2.0 and those on LogiQA 1.0 to understand the improvements we made to the datasets.

A. Settings

We re-implement the rule-based methods strictly following Yih et al. [59] for MRC. For both NLI and MRC, we follow the HuggingFace implementation [64] and use the OpenAI API, taking the off-the-shelf model BERT-base, RoBERTa-base, XLNet-base BART-base, and GPT-2 for LogiQA 2.0, and Chinese BERT-base [65], Chinese RoBERTa-base [65], Chinese XLNet-base [66], Chinese BART-base [67] and Chinese GPT-2 [42] for the Chinese LogiQA 2.0 MRC section. All models are trained for 20 epochs. We find hyper-parameters using grid search: batch size $\in \{8, 16, 32\}$, learning rate $\in \{1e-5, 2e-5, 3e-5, 4e-5, 5e-5\}$ and gradient accumulate step $\in \{1, 2, 4\}$. We set the max length to 256 tokens for all models. All hyper-parameters are decided by the model performance on the development sets.

B. NLI results

Table III shows the results of models discussed in Section III. Following the NLI benchmark-setting [46] [11], we employ the overall accuracy as the main evaluation method. Furthermore, to give a more detailed analysis, we also calculate the precision (P), recall (R), and F1-score (F1). The F1-score of human performance on the test set is 91.40%, and the human accuracy performance is 89.36% for LogiQA 2.0; BERT-base yields 54.87% accuracy, and the F1-score is 59.61%; RoBERTa-base gives slightly better results with 55.43% accuracy and 57.96% F1-score. XLNet, BART and GPT-2 give similar results. The GPT-3 model gives the best performance with 67.78% accuracy and 68.65% F1-score. This shows that pre-training is effective for the task, and the contrast between GPT-2 and GPT-3 shows the effectiveness of increased model size. In addition, there is still a significant gap between the performance of pre-trained language models and human performance. The result concurs with the work of Clark et al. [68]. That is, although yes/no questions frequently occur in real-world scenarios, binary classification is notoriously hard for language models [68].

We rerun a series of experiments of BERT-base model fine-tuning on several NLI benchmarks, and the results are shown in Table II. Each experiment is trained for 10 epochs with the same hyperparameters. As seen from Table II and Table III, compared to their performance on traditional NLI benchmarks and human performance, there is still a gap, which shows that LogiQA 2.0 is challenging for the NLU models.

C. MRC results

Table IV shows the results of the models discussed in Section IV. In particular, the human performance is 84.00%, and the ceiling performance is 98.00% for LogiQA 2.0, which shows that the difficulty level of the dataset is not high for human testees. In contrast, all algorithmic models perform significantly worse than humans, demonstrating that the methods are relatively weak in logical-reasoning intensive reading comprehension. In addition, results on the Chinese dataset are on the same level as those on the English dataset.

In particular, the rule-based methods give accuracies of 30.25% and 31.17% on the development set and test set, respectively, slightly higher than a random guess baseline. This shows that using lexical matching alone makes the questions extremely difficult to solve. The first problem in Figure 3(b) serves as one intuitive example. Option C of the problem has the least lexical overlapping yet is the right answer; the lexical matching method fails in such cases.

On LogiQA 2.0, such models perform better than the methods without contextualized embeddings. It has been shown that pre-trained models have a certain degree of commonsense and logical capabilities [69]. However, the best result by GPT-3 is 56.21%, still much below human performance. This shows that knowledge in pre-trained models is rather weak for logical reasoning. It remains an open question how deep learning machine readers can be equipped with solid reasoning capability. In addition, the better results on NLI in Table III as compared to MRC in Table IV by the same models result from the fact that NLI is a binary classification task, which offers a playground for testing logical reasoning capabilities with fewer answer-level ambiguities.

LogiQA 2.0 versus LogiQA 1.0. We nearly double the data size from the LogiQA 1.0 to the LogiQA 2.0 MRC section. To examine how the MRC data size affects the performance of pre-trained models, we compare the performance of the same data portion of LogiQA 2.0 as in LogiQA 1.0 versus the Full LogiQA 2.0. As shown in Table IV and V, the RoBERTa-base result on the test set of the whole data is 45.42% accuracy, significantly higher than that of the partial data, which is 36.56% accuracy. We see a significant performance improvement on the LogiQA 2.0 MRC task for state-of-the-art MRC models, which happens due to the enhanced data quality and increased data size.

The influence of English translation quality on LogiQA 2.0. We reran experiments using the same data as LogiQA 1.0, but with the new English translation. We use the same training set for training and the same dev and test set for testing. The results are shown in the “2.0 portion” column Table V. On the dev set, the accuracy of BERT improved from 33.83% to 36.75%, RoBERTa gives an accuracy of 36.98%, higher than on LogiQA 1.0. Similarly, the accuracy of the other models also improved. This shows that the quality of English translation has a positive influence on the MRC result.
Fig. 6: RoBERTa results on the ablation tests (accuracy%).

VI. ANALYSIS

We give a detailed analysis based on the empirical results of both NLI and MRC models on the LogiQA 2.0 test sets.

A. Ablation Tests

Following recent studies [70] [71], we conduct a set of ablation experiments using RoBERTa to measure bias in the dataset by checking the performance based on partial information. To verify the data quality against bias and artifacts [72] For the NLI ablation test, we report the hypothesis-only and context-only results of the RoBERTa model. For the MRC ablation test, we report the paragraph & option result, the question & option result, and the option-only result for the RoBERTa model. In addition, the NLI and MRC tasks’ random results are theoretically given, respectively. Figure 6 shows the results.

The LogiQA 2.0 NLI task. As we can see from Figure 6, training on context-only data yields 47.15% accuracy, and training RoBERTa on hypothesis-only data yields 50.92% accuracy. The results show that the RoBERTa model cannot decide the entailment relation between premise and hypothesis, given only one party. This shows that LogiQA 2.0 NLI does not have noticeable data artefacts [71].

The LogiQA 2.0 MRC task. As can be seen, the results of the MRC ablation test show that there is a significant drop in accuracy without the paragraph, the question, or both, which indicates that the bias on the dataset is weak. In particular, without the input paragraph, the accuracy drops from 45.42% to 29.47%; If we omit the question, the accuracy of the RoBERTa model is 35.37%, still 10% lower than the performance of RoBERTa on the complete data; The accuracy of the RoBERTa model on option-only data is 28.06%, slightly higher than a random guess. The ablating question causes a relatively minor performance drop as compared with the paragraph, which is consistent with observations by Huang et al. [69]. This is likely because the diversity of questions is lower than the paragraph. The above results show that our dataset does not have a strong bias.

B. Transfer Learning

Recent studies have shown the benefit of fine-tuning models on similar datasets for knowledge transfer [69]. The set of experiments below aims to understand the underlying correlation and contrasts between the NLI and MRC sections in LogiQA and relevant existing NLI and MRC benchmarks.

Knowledge transfer from MNLI to the LogiQA 2.0 NLI task. For the LogiQA 2.0 NLI task, we explore the MNLI [47] dataset for knowledge transfer. In particular, we first train RoBERTa with MNLI training data (RoBERTa-NLI), and then we use the trained model for fine-tuning the LogiQA 2.0 NLI data. As shown in the second rows of Table VI(a), RoBERTa-NLI only achieves 47.18% on the test set of LogiQA 2.0, much lower than 55.43% (Table III) when trained on LogiQA 2.0, and even lower than the theoretical random baseline, which shows that features learned over MNLI cannot directly transfer to LogiQA. After fine-tuning models on the NLI section of LogiQA 2.0, BART-NLI-FT gives a better performance of 57.44% which is slightly better than training only on LogiQA 2.0, which shows that LogiQA is unique compared to traditional NLI benchmarks because the reasoning types it needs are drastically different.

Knowledge transfer from RACE/COSMOS to the LogiQA 2.0 MRC task. Similarly, for the MRC section of the LogiQA 2.0, we conduct a set of transfer learning experiments to understand the degree of overlap in terms of necessary knowledge for solving problems in our dataset and existing datasets. In particular, we first fine-tune the RoBERTa model on a source dataset before fine-tuning the model on LogiQA 2.0. The model performance is expected to increase if the required knowledge is similar. RACE and COSMOS are adopted as the source datasets. The former tests English reading skills, while the latter tests commonsense knowledge. As shown in Table VI(b), the RoBERTa model trained only on either source dataset gives significantly lower accuracies on the LogiQA 2.0 test set compared with the RoBERTa model trained on LogiQA 2.0. The performance of RoBERTa trained on RACE is even close to the random guess baseline. In addition, further fine-tuning on LogiQA 2.0 leads to improvements over the source-trained baselines, but the resulting models do not outperform a model trained only on LogiQA 2.0. The observation is different from most other datasets [69], [73], demonstrating that LogiQA 2.0 contains highly different challenges compared with existing datasets.
Transfer learning from the LogiQA 2.0 NLI task to the LogiQA 2.0 MRC task. Natural language inference for downstream tasks is explored in recent research [74] [75] [76]. NLI models are used to verify the correctness of QA systems’ prediction because logical entailment provides more information than selecting the highest score from multi-choice options. [75]. We conduct a transfer learning experiment across the LogiQA 2.0 NLI and the LogiQA 2.0 MRC tasks, using RoBERTa-base for our experiment. The result is shown in Figure VI(b). The RoBERTa model is first trained on the LogiQA 2.0 NLI section. Then we use the LogiQA 2.0 MRC section to fine-tune the RoBERTa model trained on the LogiQA 2.0 NLI section, which yields an accuracy of 48.28% on the test set, the highest score of the RoBERTa model. The results show that a model trained on a natural language inference dataset can improve the downstream tasks like question answering after fine-tuning.

C. Performance Across Different Input Lengths

We measure the accuracy of RoBERTa against the input size for both NLI and MRC. In particular, for the NLI task, we concatenate the premise and hypothesis to calculate the context length for each instance; for the MRC task, the number of words in the paragraph, the question, and the candidate answers are added together as the length of a test instance. The results are all shown in Figure 7, where the model performances are not negatively associated with the input size, which is different from most NLP benchmarks [77]. This shows that the level of challenge in logical reasoning can be independent of the input verbosity for both NLI and MRC.

D. Lexical Overlap

Lexical overlap between input and output components has been shown to lead to typical spurious features for various NLP tasks [78] For the NLI task, we examine the overlap ratio between premises and hypotheses to understand the surface matching bias of the NLI data section. For the MRC task, we aim to understand the bias of models in selecting the candidate answers that have the best surface matching with the paragraph. To this end, we calculate the unigram overlap between each candidate answer and the given paragraph for each problem and mark the best-matching candidate.

We report the “Overlap Ratio” by calculating the accuracy between model prediction and the best-matching candidate. The results are shown in Table VII. For NLI, Table VII(a) shows that the gold-standard output of the NLI data has an accuracy of 29.04%, lower than that of pre-trained models, which indicates a certain level of surface matching by the models. As seen in Table VII(b), the gold-standard output of the MRC data has an accuracy of 27.85%, whilst all of the models give accuracies above this number, which shows a tendency of superficial matching. In particular, RoBERTa gives a lower matching accuracy, showing that it relies the least on lexical patterns than BERT. As can also be seen from the tables, surface patterns can be a spurious feature, leading to lower results.
TABLE VIII: Test accuracy (%) per reasoning type.

<table>
<thead>
<tr>
<th>Reasoning Type</th>
<th>NLI section</th>
<th>MRC section</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Ratio</td>
</tr>
<tr>
<td>Categorical reasoning</td>
<td>78.06</td>
<td>43.09%</td>
</tr>
<tr>
<td>Sufficient conditional reasoning</td>
<td>55.92</td>
<td>21.40%</td>
</tr>
<tr>
<td>Necessary conditional reasoning</td>
<td>49.11</td>
<td>22.06%</td>
</tr>
<tr>
<td>Disjunctive reasoning</td>
<td>47.15</td>
<td>16.87%</td>
</tr>
<tr>
<td>Conjunctive reasoning</td>
<td>49.58</td>
<td>31.63%</td>
</tr>
</tbody>
</table>

E. Reasoning Types

Table VIII gives the performances of RoBERTa over the 5 reasoning types discussed in Section II-C. For both NLI and MRC, the method gives the best accuracy on categorical reasoning. However, the results are significantly lower for the other four reasoning types. We give qualitative discussion via case study to understand why these tasks are challenging.

a) Categorical reasoning: P1 of Figure 8 shows a typical example where the definition of children’s products is given in the paragraph, and the testee is asked to select a correct instance. A key here is the age range (i.e., under 12). RoBERTa incorrectly chooses the candidate that is superficially similar to the paragraph while ignoring the reasoning process.

b) Conditional reasoning: P2 of Figure 8 is a representative example of the most challenging conditional reasoning questions. In particular, a variety of sufficient and necessary conditional relations are given in the paragraph, which includes:

\[ x = \text{“Salmons swim”} \]
\[ y = \text{“Sea eagles and brown bears gather”} \]
\[ z = \text{“Hydropower dam exists downstream”} \]
\[ w = \text{“Trees lose leaves”} \]
\[ x \rightarrow w \text{ (Necessary conditional relation)} \]
\[ y \rightarrow x \text{ (Sufficient conditional relation)} \]
\[ x \rightarrow \neg z \text{ (Sufficient conditional relation)} \]

The correct answer depends on fully understanding both the necessary and sufficient conditional reasoning facts. RoBERTa makes a mistake by ignoring the “not” operator in the \[ x \rightarrow \neg z \] condition, which coincides with prior observations on BERT and negation [79].

c) Conjunctive and disjunctive reasoning: P3 of Figure 8 represents one of the most challenging questions in the dataset, where the premises and candidate give a set of constraints in both conjunctive and disjunctive forms, and the question asks which candidate conforms to the premises. The testee is expected to enumerate different possible situations and then match the cases to the candidates by thoroughly understanding the candidates also. Intuitively, RoBERTa is not directly equipped with such reasoning capacity.

In summary, the observations above indicate that RoBERTa can rely on relatively superficial patterns to resolve ambiguities for both NLI and MRC, rather than learning to use formal logical reasoning. As a result, it learns spurious features [80]–[82], which can lead to accuracies below random guess baselines when the training and testing distributions consistently differ.

TABLE IX: Models performance on LogiQA 2.0 NLI test set w/o major/minor premise annotation. (accuracy%)

F. The Role of Major/Minor Premise Annotation

The optional major and minor premises (Section II) are potentially useful for logical reasoning in NLI. We conduct experiments to examine the role of major/minor premise annotation. As mentioned in Section III-A, special tokens are added to the beginning of both major and minor premises. We use the same hyperparameters as the main NLI experiments. The results are shown in Table IX. While BART and GPT-3 give better accuracy and F1-score with the major/minor premise indicators, results of BERT-RoBERTa and GPT-2 are relatively less improved. Overall, major and minor premises show a positive influence on model performance. We leave a further investigation of how to effectively make use of such labels to future work.

VII. RELATED WORK

A contrast between LogiQA 2.0 and related datasets is shown in Table X. To our knowledge, LogiQA is the first dataset to investigate different types of formal logical reasoning in both MRC and NLI tasks.

Datasets for Traditional Natural Language Inference. Different schemes of entailment relations are investigated in NLI research, where two-way [96]–[98] and three-way [46], [47] classification are most frequently used in NLI datasets. The task of Natural Language Inference was first introduced as Recognizing Textual Entailment (RTE-1) [96]. It classifies sentence pairs into either entailment or non-entailment, a two-way classification scheme. The advantage of binary classification is that it avoids the vagueness between neutral and contradiction, which undermines the annotation quality for crowd-sourcing datasets. Similar to the RTE dataset, we use binary classification in the NLI version of LogiQA 2.0. However, the RTE dataset is different from our dataset in that the context of the dataset is only a single sentence, and it is not designed for logical reasoning.

The FraCas test suite [53] is hand-crafted by experts with hundreds of NLI questions. Similar to our dataset, it covers logic phenomena like quantifiers and comparatives. However, unlike our dataset, it is not designed for complex logical reasoning with multiple sentence inputs. In addition, the dataset is rather small and thus more suitable for validating rule-based methods. The first large-scale NLI dataset that can be used for the training of neural models is SICK [89]. Different from our work, the SICK dataset is crowd-sourced and has unreliability issues being reported [99], and is not widely used in training neural models. Finally, SICK introduces a three-way NLI scheme instead of a two-way classification.
For neural models, the most popular NLI datasets for NLU benchmarks are the SNLI [46] and the Multi-genre NLI (MNLI) [11] datasets, which are inspired by the SICK dataset. Compared to SICK, the two datasets improve in data scale, label quality, and also coverage of genres. They catalyze the development of models with better representation abilities and become the go-to benchmark for natural language understanding research. Through SNLI and MNLI are widely used in NLP research, researchers find that they contain biases that can be exploited by models to achieve high scores without the real inference abilities [78] [100]. Datasets generated through crowd-sourcing are prone to annotation artefacts [72].

On the contrary, our dataset is sourced from expert-designed logical reasoning tests, making it more reliable in data quality. ENTAILMENTBANK [90] is a dataset for multistep, multi-premise textual entailment in the form of question-answering. The dataset is derived from grade-school-level multi-choice science questions. Annotators are instructed to construct entailment trees given a hypothesis (question-answer pair) and relevant text. With the hypothesis and its context from each instance, the dataset can be used for the NLI task. Similar to our dataset, the context of ENTAILMENTBANK is beyond the sentence level. However, it does not explore logical reasoning, which is the main focus of our dataset.

Datasets for Logical Reasoning NLI. RuleTaker [25] is a synthetic dataset to examine the reasoning ability of transformers [62] over natural language rules. With input facts and input rules as context, the output is a binary true-or-false answer. Although originally designed for question-answering, the dataset can be easily converted into NLI-style. Different from the expert-designed LogiQA 2.0, the dataset is in synthetic English following templates. MED [91] and HELP [92] are two NLI datasets focusing on monotonicity reasoning, which is an essential concept in Natural Logic [101]. The datasets are generated through monotonicity rules and only investigate monotonicity-related inference specifically. Similar to the NLI section of our dataset, MED and HELP investigate basic logic phenomena in natural language, which is monotonicity in particular. However, unlike their work, our dataset is human written, and tackles more complex formal logical reasoning over multiple sentences.

ConjNLI [94] is a challenging stress test for NLI over conjunctive sentences, where the premise differs from the hypothesis by having conjuncts being removed, added, or replaced. Similar to LogiQA 2.0, Logical reasoning about conjunctions is heavily tested. However, our dataset covers broad logical reasoning types. Moreover, in ConjNLI premise-hypothesis pairs are created automatically by applying conjunct operations on collected conjunctive sentences. Different from their work, our NLI version of LogiQA is converted from question-answering tests, which is more diverse.

TaxiNLI [95] is an NLI dataset re-annotated on the MNLI dataset with fine-grained category labels. The annotation includes logical categories like connectives, mathematical, and deduction, which is similar to our dataset. However, unlike our dataset, TaxiNLI contains only a small fraction of logical reasoning examples and is not designed particularly for logical reasoning. ConTRoL [44] is an NLI dataset that further

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Logic</th>
<th>Context</th>
<th>Source</th>
<th>Size</th>
<th>Schemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD [12]</td>
<td>Reading Comprehension</td>
<td>✗</td>
<td>Passage</td>
<td>Wikipedia</td>
<td>100K</td>
<td>Question Answering</td>
</tr>
<tr>
<td>TriviaQA [83]</td>
<td>Reading Comprehension</td>
<td>✗</td>
<td>Passage</td>
<td>Trivia Websites</td>
<td>95K</td>
<td>Question Answering</td>
</tr>
<tr>
<td>DuoRC [84]</td>
<td>Reading Comprehension</td>
<td>✗</td>
<td>Passage</td>
<td>Movie</td>
<td>186K</td>
<td>Question Answering</td>
</tr>
<tr>
<td>NarrativeQA [85]</td>
<td>Reading Comprehension</td>
<td>✗</td>
<td>Passage</td>
<td>Movie</td>
<td>46K</td>
<td>Question Answering</td>
</tr>
<tr>
<td>DROP [86]</td>
<td>Reading Comprehension</td>
<td>✗</td>
<td>Passage</td>
<td>Wikipedia</td>
<td>96K</td>
<td>Question Answering</td>
</tr>
<tr>
<td>COSMOS QA [69]</td>
<td>Reading Comprehension</td>
<td>✗</td>
<td>Passage</td>
<td>Webblog</td>
<td>35588</td>
<td>Question Answering</td>
</tr>
<tr>
<td>MuTual [87]</td>
<td>Next Utterance Prediction</td>
<td>✗</td>
<td>Dialogue</td>
<td>Exam</td>
<td>8860</td>
<td>Multi-choice Selection</td>
</tr>
<tr>
<td>ReClor [88]</td>
<td>Reading Comprehension</td>
<td>✓</td>
<td>Passage</td>
<td>Exam</td>
<td>6138</td>
<td>Question Answering</td>
</tr>
<tr>
<td>LogiQA [34]</td>
<td>Reading Comprehension</td>
<td>✓</td>
<td>Passage</td>
<td>Exam</td>
<td>8678</td>
<td>Question Answering</td>
</tr>
<tr>
<td>LogiQA 2.0 (MRC)</td>
<td>Reading Comprehension</td>
<td>✓</td>
<td>Multi-Sentence</td>
<td>Exam</td>
<td>16K</td>
<td>Question Answering</td>
</tr>
<tr>
<td>FraCaS [53]</td>
<td>Natural Language Inference</td>
<td>✓</td>
<td>One-Sentence</td>
<td>Diverse</td>
<td>346</td>
<td>Question Answering</td>
</tr>
<tr>
<td>RTE (GLUE) [57]</td>
<td>Natural Language Inference</td>
<td>✗</td>
<td>One-Sentence</td>
<td>Diverse</td>
<td>5K</td>
<td>Two-way Classification</td>
</tr>
<tr>
<td>SICK [89]</td>
<td>Natural Language Inference</td>
<td>✗</td>
<td>One-Sentence</td>
<td>Captioning</td>
<td>10K</td>
<td>Three-way Classification</td>
</tr>
<tr>
<td>SNLI [46]</td>
<td>Natural Language Inference</td>
<td>✓</td>
<td>One-Sentence</td>
<td>Captioning</td>
<td>570K</td>
<td>Three-way Classification</td>
</tr>
<tr>
<td>ENTAILMENTBANK [90]</td>
<td>Natural Language Inference</td>
<td>✓</td>
<td>Passage</td>
<td>Exam</td>
<td>1840</td>
<td>Question Answering</td>
</tr>
<tr>
<td>RuleTaker [25]</td>
<td>Natural Language Inference</td>
<td>✓</td>
<td>Passage</td>
<td>Synthetic</td>
<td>800K</td>
<td>Question Answering</td>
</tr>
<tr>
<td>MED [91]</td>
<td>Natural Language Inference</td>
<td>✗</td>
<td>One-Sentence</td>
<td>Crowd-sourcing</td>
<td>5382</td>
<td>Two-way Classification</td>
</tr>
<tr>
<td>HELP [92]</td>
<td>Natural Language Inference</td>
<td>✓</td>
<td>One-Sentence</td>
<td>PMB [93]</td>
<td>36K</td>
<td>Two-way Classification</td>
</tr>
<tr>
<td>ConjNLI [94]</td>
<td>Natural Language Inference</td>
<td>✗</td>
<td>One-Sentence</td>
<td>Wikipedia</td>
<td>1623</td>
<td>Three-way Classification</td>
</tr>
<tr>
<td>TaxiNLI [95]</td>
<td>Natural Language Inference</td>
<td>✗</td>
<td>One-Sentence</td>
<td>MNLI</td>
<td>10K</td>
<td>Three-way Classification</td>
</tr>
<tr>
<td>ConTRoL [44]</td>
<td>Natural Language Inference</td>
<td>✓</td>
<td>Passage</td>
<td>Exam</td>
<td>8325</td>
<td>Three-way Classification</td>
</tr>
<tr>
<td>LogiQA 2.0 (NLI)</td>
<td>Natural Language Inference</td>
<td>✓</td>
<td>Multi-Sentence</td>
<td>Exam</td>
<td>35K</td>
<td>Two-way Classification</td>
</tr>
</tbody>
</table>

TABLE X: Comparison between our dataset and existing benchmarks of MRC and NLI. (✓ indicates datasets addressing logical reasoning, ✗ indicates datasets not particular for logical reasoning.)
investigates contextual reasoning under the NLI framework. It has 36.2% of premise-hypothesis pairs under the category of logical reasoning. Similarly, it has a portion of instances addressing multi-sentence logical reasoning. However, it is not an NLI dataset solely for formal logical reasoning but for contextual reasoning in general. Therefore, to enlarge the NLI version of LogiQA 2.0, we incorporate the logical reasoning instances into our new NLI dataset.

Datasets for Traditional Machine Reading Comprehension. A seminal dataset for large-scale reading comprehension is SQuAD [12], which requires selecting a factual answer from all possible spans in a given passage. Many neural methods have been developed for this dataset, achieving results that rival human testees. As a consequence, more reading comprehension datasets with increasing challenges are proposed. These datasets can be classified according to the main challenges. In particular, TriviaQA [83] requires evidence integration across multiple supporting documents to answer the questions. DuoRC [84] and Narrative QA [85] raise challenges by introducing two passages about the same facts. Welbl et al. [102] and HotpotQA [35] test models for text understanding with sequential multi-step reasoning. Drop [86] tests discrete numerical reasoning over the context. MuTual [87] tests dialogue reasoning ability via the next utterance prediction task. The answer (or candidate in multi-choice-questions) is mostly a text span in the given passage for the above datasets. Several types of reasoning are necessary, such as geolocational reasoning and numerical computation. Unlike these datasets, our dataset contains answers not directly included in the input passage and requires comprehensive reasoning methods beyond text-matching-based techniques.

Similar to our dataset, recent datasets for commonsense reasoning, including MCScript [103] and COSMOS [69], also contain candidate answers not directly included in the input passage. They test a model’s capability of making use of external background knowledge about spatial relations, cause and effect, scientific facts, and social conventions. In contrast, our dataset focuses on logical reasoning, and most necessary facts are not directly included in the given passage. In addition, most of the existing datasets are labeled by crowd-sourcing. In contrast, our dataset is based on examination problems written by human experts for students and therefore has a better guarantee of quality. This is particularly important for datasets that involve abstract reasoning skills.

Datasets for Logical Reasoning Reading Comprehension. LogiQA 1.0 [34], our previous dataset, was one of the earliest datasets for testing logical MRC. Reclor [88] is another reading comprehension dataset requiring logical reasoning. It collects question-answering examples from the LSAT exams 7, which are targeted to test human logical reasoning abilities. Compared to our dataset, the size of Reclor is smaller; and the reasoning types that Reclor uses are the same as the question types of the original tests, such as “Identify a Flaw”, which is informally defined. In contrast, we annotate formally defined logic reasoning types for LogiQA and investigate formal logic.

NLI Datasets Converted from MRC Datasets. While multiple MRC datasets are available, obtaining high-quality NLI datasets is difficult [48]. Moreover, annotating large-scale NLI datasets through crowd-sourcing is costly and suffers from human artefacts [104]. Thus, converting question-answering datasets into natural language inference datasets has become a promising research direction [48] [105] [75]. This approach is awarded for the fact that abundant large-scale MRC datasets exist, making it an efficient way of creating high-quality NLI datasets. The Question-answering NLI (QNLI) dataset [57] is an NLI dataset automatically derived from the Stanford Question Answering Dataset (SQuAD), an MRC dataset we talked about in VII. The dataset was converted into sentence pair classification by forming a sentence pair between the question and each sentence in the corresponding context and filtering out pairs with low lexical overlap between the question and the context. The task is to determine whether the context sentence contains the answer to the question. Similar to our dataset, QNLI is derived from question-answering datasets, and the framework is also a two-way classification. However, different from their work, the LogiQA 2.0 NLI section is converted from expert-designed logical reasoning questions other than Wikipedia passages; the premise of each instance is multi-sentence other than single-sentence. Similarly, QA-NLI [105] converts question-answering pairs into their declarative forms and obtains 500K NLI examples. The source datasets come from existing MRC benchmarks. Different from our dataset, QA-NLI adopts a three-way classification scheme; it is not a dataset for logical reasoning.

Other Datasets Addressing Logical Reasoning. Apart from the datasets we mentioned above, there have been existing datasets related to logical reasoning in other NLP tasks. In particular, Habernal et al. [106] designs a dataset for argument reasoning, where a claim is given, and the model is asked to choose a correct premise from two candidates to support the claim. Similar to our dataset, the dataset concerns deductive reasoning. The most significant difference between our dataset and this dataset is that they focus on computational argumentation. The form of their task is neither a reading comprehension nor a textual entailment task. In addition, our dataset has more instances (15,708 vs. 1,970), more choices per question (4 vs. 2), and is written by relevant experts rather than being crowd-sourced. CLUTTR [107] is a dataset for inductive reasoning over family relations. The input is a given passage and a query pair, and the output is a relationship between the pair. The task is reasoning on a fixed domain (i.e., family relationship), which is in line with prior work on social relation inference [108]. In contrast, our dataset investigates formal logical reasoning with various types.

VIII. CONCLUSION

We presented LogiQA 2.0, a new version of LogiQA with increased data size, refined translation, and a new NLI counterpart to the original MRC task. Experiments show that the extension leads to better performance with a variety of pre-trained models due to quality improvements and data size augmentation. In addition, comparison experiments between

---

7 https://www.lsac.org/lsat
model and human performances show that the current state-of-the-art methods still lag far behind the human level for logical NLU. To our knowledge, the NLI section of LogiQA 2.0 is the first large-scale expert-designed NLI benchmark for logical reasoning.

REFERENCES

此由事则兴上富国非强民或富民如者强则。者节辱或礼知秋⾐果实礼⻝知的春国，如时期⺠于节⾜之之。

对于春秋时期的某国，如果诸侯实或衣⾰，则民知礼节或知荣耀。如果民知礼节或知荣者，则或居者国，或者民强。如果民强，则百业兴。事实上该国并非国富，而且并非百业兴。由此可推出？
A.该国诸侯实但并非衣⾰。
B.该国并非诸侯实但衣⾰。
C.该国诸侯实且衣⾰。
D.该国并非诸侯实且并非衣⾰。

Fig. 9: Examples of items deleted from the dataset.
### Reading Comprehension

**Text:** Some Cantonese don't like chili. Therefore, some southerners don't like chili.

**Question:** Which of the following question can verify above argument?

**Options:**
- A. Some Cantonese like to eat chili.
- B. Some people who like to eat chili are southerners.
- C. All Cantonese are southerners.

**Text:** It is difficult for cactus to survive in humid climates; citrus is difficult to grow in cold climates. In most parts of a province, at least one species is not difficult to survive and grow between cactus and citrus.

**Question:** If the above assertion is true, which of the followings must be false?

**Options:**
- A. Half of the province is humid and cold.
- B. Most of the province is hot.
- C. Most of the province is either dry or warm.
- D. The climate is not cold and humid in the whole province.

### Natural Language Inference

**Major Premise:** All Cantonese are southerners.

**Minor Premise:** Some Cantonese don't like chili.

**Hypothesis:** Some southerners don't like chili.

**Major Premise:** It is difficult for cactus to survive in humid climates; citrus is difficult to grow in cold climates.

**Minor Premise:** In most parts of a province, at least one species is not difficult to survive and grow between cactus and citrus.

**Hypothesis:**
- 1. Half of the province is humid and cold.
- 2. Most of the province is hot.
- 3. Most of the province is either dry or warm.
- 4. The climate is not cold and humid in the whole province.

**Entailed:** ✓

**Not Entailed:** ✗

**Fig. 10:** Converting a multi-choice reading comprehension data into a Natural Language Inference data.