Hence, Socrates is mortal: A Benchmark for Natural Language Syllogistic Reasoning

Yongkang Wu[1], Meng Han[1], Yutao Zhu[2], Lei Li[1], Xinyu Zhang[1], Ruofei Lai[1], Xiaoguang Li[3], Yuanhang Ren[1], Zhicheng Dou[4] and Zhao Cao[1]*

[1]Huawei Poisson Lab, China
[2]University of Montreal, Montreal, Quebec, Canada
[3]Huawei Noah's Ark Lab, China
[4]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
{wuyongkang7,zhangxinyu35,caozhao1}@huawei.com

Abstract

Syllogistic reasoning, a typical form of deductive reasoning, is a critical capability widely required in natural language understanding tasks, such as text entailment and question answering. To better facilitate research on syllogistic reasoning, we develop a benchmark called SYLLOBASE that differs from existing syllogistic datasets in three aspects: (1) Covering a complete taxonomy of syllogism reasoning patterns; (2) Containing both automatically and manually constructed samples; and (3) Involving both the generation and understanding tasks. We automatically construct 50k template-based syllogism samples by mining syllogism patterns from Wikidata and ConceptNet. To improve our dataset's naturalness and challenge, we apply GPT-3 to paraphrase the template-based data and further manually rewrite 1,000 samples as the test set. State-of-the-art pre-trained language models can achieve the best generation ROUGE-L of 38.72 by T5 and the best multi-choice accuracy of 72.77% by RoBERTa on SYLLOBASE, which indicates the great challenge of learning diverse syllogistic reasoning types on SYLLOBASE. Our datasets are released at https://github.com/casually-PYlearner/SYLLOBASE.

1 Introduction

Reasoning, as a typical way for human beings to obtain new knowledge and understand the world, is also an ultimate goal of artificial intelligence (Newell and Simon, 1956; Lenat et al., 1990). Reasoning skills, *i.e.*, examine, analyze, and critically evaluate arguments as they occur in ordinary language, have been required by many natural language processing tasks, such as machine reading comprehension (Liu et al., 2020; Yu et al., 2020), open-domain question answering (Kwiatkowski et al., 2019; Huang et al., 2019), and text generation (Dinan et al., 2019).[1] According to different mental processes, reasoning can be categorized as deductive, inductive, abductive, etc (Copi et al., 2016). In Piaget's theory of cognitive development (Huitt and Hummel, 2003), these logical reasoning processes are necessary to manipulate information, which is required to use language and acquire knowledge. Therefore, the study of logical reasoning is worthy of our attention because it is so prevalent and essential in our daily lives.

In this study, we focus on syllogism, which is a typical form of reasoning and has been studied for a long time (it was initially defined in Aristotle's logical treatises *Organon*, composed around 350 BCE). As shown in Table 1, a syllogism often contains two premises and a conclusion, where the conclusion can be inferred based on the given premises through a deductive reasoning process.[2] Though reasoning-required tasks (such as question answering) have been widely studied, the thorough study to test the deductive reasoning capabilities of a model or system is rare. In the study of syllogism, there are only a few datasets, and they have several limitations: (1) They focus merely on categorical syllogism (shown in Figure 1) (Dames et al., 2020; Dong et al., 2020; Aghahadi and Talebpour, 2022). Even though it is the most common type, syllogisms come in a variety of forms. They involve different reasoning processes and are also beneficial. (2) Some datasets (Dames et al., 2020; Dong et al., 2020) are not in natural language, which are difficult to adapt to inference requirements in real natural language scenarios. (3) More severely, all of them have less than 10k samples, which are insufficient for training deep neural networks.

To support further study on syllogistic reasoning, in this work, we build a new natural language

*Corresponding author.

[1]The definition of logical reasoning, https://www.lsac.org/lsat/taking-lsat/test-format/logical-reasoning.

[2]There can also be three or more premises. More details are given in Section 3.2.4.

Findings of the Association for Computational Linguistics: ACL 2023, pages 2347–2367

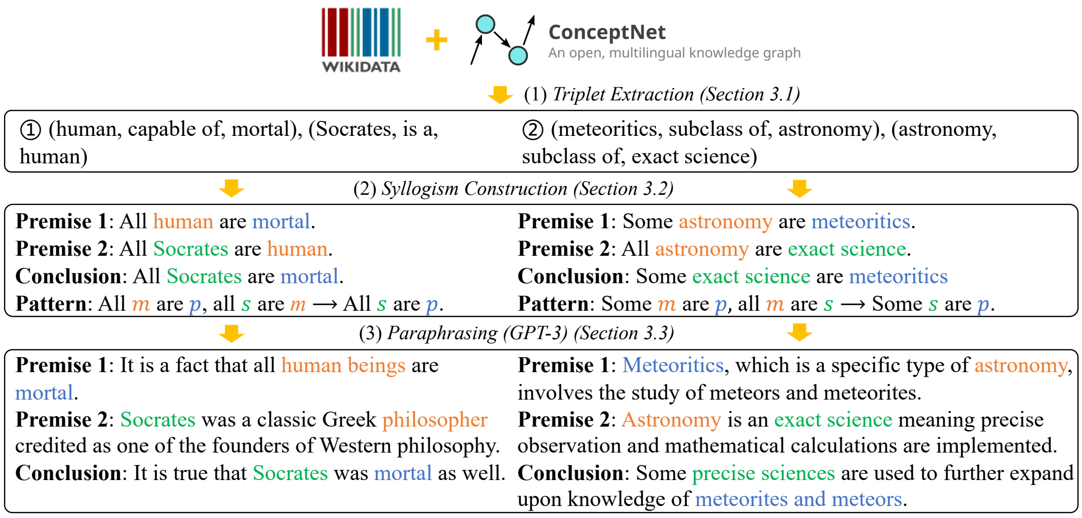

Figure 1: Illustration of our data construction process. Two examples are shown, and the colored terms correspond to the symbols in the pattern.

benchmark—SylloBase, which has the following features (some examples are shown in Table 2): **First**, it is a more complete benchmark that covers five types of syllogisms. Therefore, it can support more fine-grained research on certain types, their interrelationships, and their combined effect on other tasks. **Second**, all premises and conclusions are written in natural language. It more closely resembles real-world application settings in which natural language descriptions rather than categorized inputs are provided. In addition, the power of large-scale pre-trained language models can also be harnessed effectively. **Third**, with our proposed automatic construction process, we collect a large number of samples (50k in total). They can support the training of deep neural networks. In order to validate the performance on actual human syllogism, we also manually annotate 1,000 samples as the test set. This test set may also be used independently to assess the reasoning capability of models in a zero-/few-shot manner. **Finally**, to promote a more comprehensive investigation of syllogistic reasoning, we organize both a generation and an understanding task.

The experimental results indicate that there is a great deal of room for improvement in the syllogistic reasoning capabilities of existing models. Our additional experiments demonstrate the efficacy of transferring knowledge learned from our automatically constructed syllogism to actual human reasoning.

2 Background and Related Work

2.1 Syllogism

Syllogism is a common form of deductive reasoning. Basic syllogism can be categorized as categorical syllogism, hypothetical syllogism, and disjunctive syllogism. They can be further combined into polysyllogisms. In this section, we use the most common categorical syllogism to introduce the term and structure of syllogism. Other types of syllogism will be introduced in Section 3.

The left side of Figure 1 shows a well-known categorical syllogism about "Socrates is mortal". We can see a categorical syllogism usually contains two premises and a conclusion. A common term (*e.g.*, "human") links two premises, and the premises respectively define the relationship between "human" and "mortal" or "Socrates". The reasoning process is to draw a conclusion based on the premises. A syllogism can also be described by a pattern, as shown in the middle side of Figure 1.

2.2 Related Work

Syllogistic Reasoning Dataset Several syllogistic reasoning datasets have been introduced to promote the development of this field. CCOBRA (Dames et al., 2020) is a dataset with around 10k triplets (major premise, minor premise, conclusion). The task is formed as a single-choice question, and the ground-truth conclusion is shuffled with several distractors. ENN (Dong et al., 2020) is another similar dataset, but the syllogism is constructed from WordNet (Miller, 1995). SylloFigure (Peng et al., 2020) and Avicenna (Aghahadi

Table 1: Comparison of existing syllogism datasets. Our SYLLOBASE is the largest one covering all five types.

Dataset	#Types	Natural Language	Complete Patterns	Source	Size
CCOBRA	1 (Categorical)	✗ (Triplet)	✓	Crowdsourcing	10k
ENN	1 (Categorical)	✗ (Triplet)	✓	WordNet	7k
SylloFigure	1 (Categorical)	✓	✗	SNLI	8.6k
Avicenna	1 (Categorical)	✓	✗	Crowdsourcing	6k
SYLLOBASE (Our)	5	✓	✓	Knowledge Base & Crowdsourcing	51k

and Talebpour, 2022) are two natural language text-based syllogism reasoning datasets, but they are designed for different tasks. SylloFigure annotates the data in SNLI (Bowman et al., 2015), restores the missing premise, and transforms each syllogism into a specific figure.[3] The target is to predict the correct figure type of a syllogism. Avicenna is a crowdsourcing dataset, and the syllogism is extracted from various sources, such as books and news articles. These syllogisms are used for both natural language generation and inference tasks.

Different from existing datasets that focus only on categorical syllogism, our SYLLOBASE covers more types and patterns of syllogism and is significantly larger than existing datasets. More detailed comparisons are shown in Table 1.

Logic Reasoning in NLP There are several tasks and datasets related to logical reasoning in NLP. The task of natural language inference (NLI) (Bos and Markert, 2005; Dagan et al., 2005; MacCartney and Manning, 2009; Bowman et al., 2015; Williams et al., 2018), also known as recognizing textual entailment, requires model to classify the relationship types (*i.e.*, contradicted, neutral, and entailment) between a pair of sentences. However, this task only focuses on sentence-level logical reasoning, and the relationships are constrained to only a few types. Another NLP task related to logical reasoning is machine reading comprehension (MRC). There are several MRC datasets designed specifically for logical reasoning, such as LogiQA (Liu et al., 2020) and ReClor (Yu et al., 2020). A paragraph and a corresponding question are given, and the model is asked to select a correct answer from four options. This task requires models to conduct paragraph-level reasoning, which is much more difficult than NLI.

The above logic reasoning NLP tasks attempt to improve models' general logic reasoning capability, but they pay little attention to different types of reasoning processes, such as deductive reasoning or inductive reasoning. In this work, we study a specific form of deductive reasoning, *i.e.*, syllogism. We hope our benchmark can support more in-depth studies on the reasoning process.

[3] Figures in syllogism, https://en.wikipedia.org/wiki/Syllogism.

Table 2: Examples of syllogisms from our test set.

Categorical Syllogism
Premise 1: Carbon dioxide is a chemical compound.
Premise 2: Chemical compounds are considered pure substances.
Conclusion: Pure substances include carbon dioxide.

Hypothetical Syllogism
Premise 1: When you make progress in your project, you may want to celebrate.
Premise 2: Having a party is a good choice if you want to celebrate.
Conclusion: You may want to have a party if you achieve great progress in your project.

Disjunctive Syllogism
Premise 1: Newspapers are generally published daily or weekly.
Premise 2: Some newspapers are not published weekly.
Conclusion: Some newspapers are daily newspapers.

Polysyllogism
Premise 1: Some movies are not cartoon movies.
Premise 2: Science fiction animations belong to animated films.
Premise 3: Remake films are also films.
Conclusion: Some remakes are out of scope of science fiction cartoons.

Complex Syllogism
Premise 1: If Jack has computer skills *and* programming knowledge, he could write programs.
Premise 2: Jack cannot write computer programs, but he can use computers.
Conclusion: Jack does not have programming knowledge.

3 Data Construction

Our target is to develop a large-scale benchmark and support research on several typical kinds of syllogistic reasoning. It is straightforward to collect data through human annotation, as most existing datasets have explored (Dames et al., 2020; Aghahadi and Talebpour, 2022). However, this method is impracticable for obtaining large-scale data due to the high cost of human annotation. Therefore, we propose constructing a dataset automatically from existing knowledge bases and manually rewriting 1,000 samples as the test set.

3.1 Data Source

Inspired by existing studies (Dong et al., 2020) that collect data from knowledge bases, we choose Wikidata (Vrandecic and Krötzsch, 2014) and ConceptNet (Speer et al., 2017) as our data sources because they contain large-scale high-quality entities and relations.

Wikidata is an open-source knowledge base, serving as a central storage for all structured data from Wikimedia projects. The data model of Wikidata typically consists of two components: *items* and *properties*. Items represent things in human knowledge. Each item corresponds to a identifiable concept or object, or to an instance of a concept or object. We use entities in the top nine categories, including human, taxon, administrative territorial, architectural structure, occurrence, chemical compound, film, thoroughfare, and astronomical object.[4] Then, we use the relationship of *instance of*, *subclass of*, and *part of* to extract triplets.

ConceptNet is another open-source semantic network. It contains a large number of knowledge graphs that connect words and phrases of natural language with labeled edges (relations). Its knowledge is collected from many sources, where two entities are connected by a closed class of selected relations such as *IsA*, *UsedFor*, and *CapableOf*. We use ConceptNet to extract the descriptive attributes of the entities obtained from Wikidata. By this means, we can obtain another group of triplets, which are also used for constructing syllogism.

3.2 Data Processing

In this section, we introduce the construction process of five types of syllogism data, respectively. Some examples are shown in Table 2.

3.2.1 Categorical Syllogism

As shown in Table 1, a categorical syllogism is composed of a major premise, a minor premise, and a corresponding conclusion. We first construct premises and then use them to infer the conclusion and form syllogisms.

The premise in a categorical syllogism can be summarized as four propositions according to different quantifiers and copulas:

(1) All S are P; (2) No S are P;
(3) Some S are P; (4) Some S are not P;

where S and P are two entities. With different combinations of the four propositions, categorical syllogisms can be categorized into 24 valid patterns. The first part of Table 2 shows an example of Dimatis syllogism, which is one of the valid patterns.[5] To construct premises, we use the extracted triplets from Wikidata and ConceptNet. To obtain a proposition which contains negative relationship, we can use the *Antonym* and *DistinctFrom* relationship in ConceptNet to construct it. Taking the triplets (*chemical compound*, *subclass of*, *pure substance*) and (*chemical compound*, *Antonym*, *mixture*) as an example, we have:

(1) All chemical compounds are pure substances;
(2) No chemical compounds are mixture;
(3) Some pure substances are chemical compounds;
(4) Some pure substances are not mixture.

By this means, we can obtain various premises, which will be used for constructing syllogisms.

Considering the example in Table 2, which is a Dimatis syllogism, we first sample a triplet (*carbon dioxide*, *IsA*, *chemical compound*). Then, we use the middle term *chemical compound* to sample another triplet (*chemical compound*, *subclass of*, *pure substance*), which forms the minor premise. Finally, we can generate a conclusion based on the pattern definition. All other different patterns of syllogisms can be constructed in a similar way.

3.2.2 Hypothetical Syllogism

Similar to categorical syllogism, a hypothetical syllogism has two premises and a conclusion. The difference is that the premises have one or more hypothetical propositions. A hypothetical Syllogism has three valid patterns (the full list is in Appendix A), and we use five relations (*i.e.*, *Causes*, *HasSubevent*, *HasPrerequisite*, *MotivatedByGoal*, and *CausesDesire*) in ConceptNet to construct hypothetical propositions.

The following pattern is used as an example to illustrate the data construction process:

Premise 1: If P is true, then Q is true.
Premise 2: If Q is true, then R is true.
Conclusion: If P is true, then R is true.

Specifically, we extract a triplet pair where the tail entity of one triplet is the head entity of another triplet, *e.g.*, (*success*, *CausesDesire*, *celebrate*) and (*celebrate*, *CausesDesire*, *have a party*). This triplet pair can construct premises as *success makes*

[4]The full list and the statistics are available at: https://www.wikidata.org/wiki/Wikidata:Statistics.

[5]Other patterns can be referred to in Appendix A.

you want to celebrate, and *celebration makes you want to have a party*. Then, we can build a hypothetical syllogism according to the pattern, and the corresponding conclusion is *success makes you want to have a party*. Hypothetical syllogism with other patterns can be constructed in a similar way.

3.2.3 Disjunctive Syllogism

A disjunctive syllogism has two premises: One of them is a compound proposition, which tells that at least one proposition is true; The other premise tells that one proposition in the former premise is false. Then, we can infer another proposition in the former premise is true. For example, if P and Q are two propositions, a disjunctive syllogism can be described as:

Premise 1: P is true or Q is true;
Premise 2: P is not true;
Conclusion: Q is true.

According to whether the two propositions can be both true, a disjunctive syllogism can be categorized as compatible or incompatible.

We use ten relations in ConceptNet to construct disjunctive syllogism, where eight of them (such as *PartOF* and *HasA*) are used for compatible disjunctive syllogism, and the rest two (*i.e.*, *Antonym* and *DistinctFrom*) are used for incompatible disjunctive syllogism (all relations we used are listed in Appendix B). Here, we use the incompatible disjunctive syllogism as an example to illustrate the construction process.

We first sample a triplet for an entity, such as (*newspapers*, *CapableOf*, *come weekly*) and (*newspapers*, *CapableOf*, *come daily*). Then, we can construct a premise as *newspapers can come weekly or come daily*. Next, we obtain another premise, such as *some newspapers cannot come weekly*. Finally, we can have the conclusion as *some newspapers come daily*. In this way, we can automatically construct various disjunctive syllogisms based on the triplets in ConceptNet.

3.2.4 Polysyllogism

A polysyllogism is a combination of a series of syllogisms. It usually contains three or more premises and a conclusion. We construct polysyllogisms based on categorical syllogisms, and the construction process can be summarized as the following steps:

(1) We sample a categorical syllogism from our categorical syllogism repository (built in Section 3.2.1).

(2) According to the form of the conclusion, we can get its predicate term and subject term.

(3) We use these terms to traverse the repository and select a premise/conclusion that contains them.

(4) We use the conclusion obtained in the second step and the selected premise/conclusion in the third step as two new premises. Then, we can infer the conclusion and check if the generated syllogism follows a valid pattern.

(5) Repeat the above process, and we can obtain a series of syllogisms.

(6) We use both premises in the first syllogism and the minor premise in all other syllogisms as the premises of the polysyllogism. The conclusion is obtained from the last syllogism's conclusion. By this means, we can construct a polysyllogism.

We provide an example in the fourth row of Table 2 to illustrate the construction process.

3.2.5 Complex Syllogism

In addition to constructing the previous four types of syllogism, we investigate another new type of syllogism, which is called complex syllogism. A complex syllogism contains two premises and a conclusion, and the premises and conclusion are compound propositions, which contain one or more logical connectives (*i.e.*, *not*, *and*, *or*, and *if-then*). These logical connectives significantly increase the difficulty of the syllogism. An example of a complex syllogism is shown in the last row of Table 2. The construction steps can be summarized as:

(1) We randomly sample a pattern from hypothetical and disjunctive syllogism as a basic pattern.

(2) We replace the simple propositions in the basic pattern (such as P, Q, and R) by a compound proposition with the logical connectives *not*, *and*, and *or*, (*e.g.*, *not P*, *P or Q*, and *P and Q*).

(3) After the replacement, we can infer the conclusion (according to the pattern we derived, as shown in Appendix A) and construct a complex syllogism.

Rule of Replacement To replace a simple proposition by a compound proposition, we use the *Synonyms* relation in ConceptNet. For example, considering the proposition *something that might happen as a consequence of eating ice cream is pleasure*, we use the synonym of the entity *ice cream*, *i.e.*, *cone*, and construct a compound proposition as *something that might happen as a consequence of eating ice cream and cone is pleasure*.

3.3 Rewriting

With the above process, we obtain a large number of syllogisms. However, these syllogisms are constructed based on predefined patterns, which have fixed structures and may contain grammar faults. In our preliminary study, we find that models trained on such pattern-based data have a poor robustness, potentially because the models are overfitting to the patterns rather than learning the real reasoning process. To alleviate this problem, we apply GPT-3 (Brown et al., 2020) for rewriting, which has been shown to be effective (Ding et al., 2022). Specifically, we use a prompt with some human-rewritten examples to ask GPT-3 to change the expression of the syllogism but keep its original meaning and pattern. The generated results have good quality in fluency, diversity, and logic, which are suitable for training models (some examples are shown in the bottom of Figure 1, and the detailed process is described in Appendix C).

Furthermore, to test the models' performance on (real) syllogisms and facilitate future in-depth research, we manually rewrite 1,000 samples from our collected data as a test set. The rewriting process includes filtering the noise, correcting the grammar faults, and paraphrasing (details process is described in Appendix D). Our experiments (see Section 4.4) will show that the test data are very challenging, whereas training on our automatically collected data is still effective.

As yet, we have obtained 50k samples by GPT-3 rewriting, which are used for training and validation, and 1k samples by further human annotation, which are used for testing. All of them are **equally distributed** over the five types.

4 Experiments

4.1 Task Formalization

Based on our collected data, we design two tasks:

Conclusion Generation It is a natural language generation task. The model should generate the correct conclusion based on two given premises. Premises and conclusions are natural language text, which can be represented as sequences of tokens. Formally, given two premises $P_1 = \{w_1^{P_1}, \cdots, w_m^{P_1}\}$ and $P_2 = \{w_1^{P_2}, \cdots, w_n^{P_2}\}$, the model is asked to generate the conclusion $C = \{w_1^C, \cdots, w_l^C\}$, where w is a token. Similar to other text generation tasks, the generation probability of the conclusion is determined by the product of the probability of each word, which can be described as: $P(C|P_1, P_2) = \prod P(w_i^C | w_{<i}^C, [P_1; P_2])$, where [;] is concatenation operation. More premises can be handled by concatenating all of them as a long sequence.

Conclusion Selection It is a natural language understanding task. The model is asked to select a correct conclusion from four options, where three of them are distractors. Detailed construction process is given in Appendix F. With the above notations of premises and conclusion, we can define the conclusion selection task as:

$$S(C_i, [P_1; P_2]) = \frac{\exp(M(C_i, [P_1; P_2]))}{\sum_{j=1}^{4} \exp(M(C_j, [P_1; P_2]))},$$

where $S(C_i, [P_1; P_2])$ is the predicted probability of C_i as a correct conclusion, and $M(\cdot, \cdot)$ is the output logit of the model.

The statistics of our dataset for both tasks are given in Appendix G.

4.2 Baseline and Evaluation Metrics

We compare the performance of several models. For the conclusion generation task, we consider Transformer (Vaswani et al., 2017) and several pre-trained models, including GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020), and BART (Lewis et al., 2020). For the conclusion selection task, we employ BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and ELECTRA (Clark et al., 2020) as baseline methods. For all pre-trained models, we use the base version.

As for evaluation metrics, following previous studies (Aghahadi and Talebpour, 2022), we use ROUGE-1/2/L (Lin, 2004), BLEU-1/2 (Papineni et al., 2002), and BERT-Score (Zhang et al., 2020) to evaluate the performance of the conclusion generation task. ROUGE and BLEU are commonly used metrics for text generation, and they measure the n-grams overlap between the generated text and the ground-truth text. BERT-Score is a recently proposed model-based metric. It leverages the pre-trained contextual embeddings from BERT and matches words in generated and ground-truth texts by cosine similarity. For the conclusion selection task, we use Accuracy to evaluate the models' performance. The implementation details are provided in Appendix H.

Table 3: Results of conclusion generation task. "R-1/2/L" stands for Rouge-1/2/L, "B-1/2" stands for BLEU-1/2, and "BS" denotes BERT-Score.

	R-1	R-2	R-L	B-1	B-2	BS	R-1	R-2	R-L	B-1	B-2	BS	R-1	R-2	R-L	B-1	B-2	BS
Model	Categorical						Hypothetical						Disjunctive					
Transformer	15.75	2.80	14.32	5.76	0.92	82.44	19.39	2.76	18.16	12.83	2.23	86.31	18.03	2.24	16.68	8.67	1.13	83.93
GPT-2	30.98	7.07	26.12	19.86	4.26	88.22	27.93	6.65	25.38	18.54	3.93	89.63	36.68	13.32	34.83	26.21	7.51	90.72
T5	**39.03**	**11.45**	**29.55**	**23.26**	**6.43**	**89.15**	34.45	12.37	31.77	**24.71**	8.49	90.20	**50.11**	**27.67**	**47.14**	**37.55**	18.44	92.43
BART	35.19	8.88	26.86	20.73	4.18	88.93	**34.77**	**13.03**	**32.22**	24.21	**9.27**	**90.22**	49.07	27.10	46.14	36.36	**18.56**	**92.52**
Model	Polysyllogism						Complex						All					
Transformer	22.05	5.56	19.13	8.00	1.78	83.69	17.42	2.38	16.89	8.04	1.01	85.28	22.29	4.53	20.32	14.23	2.74	86.28
GPT-2	41.28	16.22	36.37	28.26	9.02	89.40	31.68	10.62	30.51	23.89	6.33	89.79	34.38	11.07	30.42	23.24	6.58	89.52
T5	45.61	20.15	40.46	**34.27**	**14.02**	90.21	**42.65**	**21.58**	**40.75**	**35.93**	**17.29**	**91.12**	**43.21**	**19.13**	**38.72**	**31.02**	**13.01**	**90.82**
BART	**46.50**	**21.18**	**41.15**	33.42	12.91	**90.37**	41.96	20.63	39.58	33.69	15.77	90.91	41.85	18.20	37.59	29.09	11.83	90.69

Table 4: Accuracy of conclusion selection task.

Type	BERT	RoBERTa	XLNet	ELECTRA
Categorical	27.50	33.00	35.00	**36.50**
Hypothetical	69.12	75.00	73.53	**77.94**
Disjunctive	97.51	97.51	**98.01**	97.51
Polysyllogism	65.02	67.49	66.50	**76.35**
Complex	68.32	70.79	71.78	**72.28**
All	64.06	**72.77**	72.67	70.89

4.3 Experimental Results

The results of all models on the conclusion generation task are shown in Table 3, while those on the conclusion selection task are reported in Table 4.

For the conclusion generation task, we can see that the overall performance in terms of word-overlap metrics (such as ROUGE and BLEU) is poor. Given that conclusions are often brief (11.84 tokens on average), these results show that the task is fairly challenging. In contrast, the BERT-Score is high, indicating that models are able to generate some semantically correct contents but cannot organize them into a reasonable conclusion. Furthermore, the pre-trained language models perform significantly better than the vanilla Transformer. We attribute this to the natural language nature of our dataset, and these results suggest that our dataset can help future research on leveraging pre-trained language models to generate logically reasonable texts. Finally, we notice that the performance on the human-written test set and the automatically generated validation set (in Table 15) is close, reflecting the good quality of GPT-3 rewriting.

For the conclusion selection task, the overall accuracy is around 70%, showing a significant deviation from perfection. In Table 4, the model for a single type of syllogism is trained solely on the corresponding type of data. Therefore, the result of type "All" is not the *average* result of the five types of syllogisms. We notice that almost all results for ELECTRA are highest, but it has only 70.89 for the type "ALL". We speculate the reason is that the ELECTRA model is not robust when trained with mixed data, and the data in different types of syllogism might confuse it. Intriguingly, the performance on categorical syllogisms is extremely bad. A potential reason is that this type of syllogisms contains more patterns (*e.g.*, categorical syllogisms have 24 valid patterns). As a comparison, the performance on hypothetical syllogisms is significantly higher since there are only three patterns. We also notice that the performance on polysyllogisms is higher than that on categorical syllogisms, despite the fact that the former is derived from the latter. We speculate the reason is that the polysyllogisms have more abundant information in premises (*i.e.*, multiple premises), which is helpful for pre-trained language models to conduct reasoning.

4.4 Further Analysis

We also explore the following research questions. To save space, we report the results of the conclusion generation task, while similar trends can be observed on the conclusion selection task, which is shown in Appendix.

Effect of Automatically Constructed Data In our benchmark, the training data are automatically constructed from knowledge bases, while the test data are human annotated.[6] To reveal the relationship between them, we conduct an additional experiment: we split the test set as new training, vali-

[6]We also perform a human evaluation on 100 automatically constructed samples (20 for each type of syllogisms). About 73% samples are grammatically perfect and logically correct. More details can be referred to at Appendix E.

Table 5: Results (ROUGE-1/2/L) of the conclusion generation task with or without pre-training on automatic training data.

Model	*w/o* Automatic data	*w/* Automatic data
GPT-2	35.35 / 11.42 / 31.75	42.39 / 15.92 / 38.25
T5	39.24 / 17.32 / 34.10	53.47 / 26.30 / 48.37
BART	42.49 / 18.41 / 38.76	50.61 / 25.51 / 47.26

Table 6: Results of transfer learning. Results in **bold** indicate improvement over non-transfer learning.

ID	Pre-training →	Fine-tuning	R-1	R-2	R-L
(1)	Categorical	Hypothetical	34.36	12.92	31.53
(2)	Categorical	Disjunctive	48.92	**27.19**	45.93
(3)	Categorical	Polysyllogism	**48.17**	**23.09**	**43.00**
(4)	Categorical	Complex	**43.95**	**22.46**	**42.14**
(5)	Polysyllogism	Categorical	**38.20**	**10.76**	**28.00**
(6)	Disjunctive	Complex	**42.77**	**21.34**	**40.42**
(7)	Hypothetical	Complex	**43.09**	**21.50**	**40.66**
(8)	Complex	Disjunctive	**49.53**	**28.51**	**47.10**
(9)	Complex	Hypothetical	34.44	11.98	31.68
(10)	None	Categorical	35.19	8.88	26.86
(11)	None	Hypothetical	34.77	13.03	32.22
(12)	None	Disjunctive	49.07	27.10	46.14
(13)	None	Polysyllogism	46.00	21.18	41.15
(14)	None	Complex	41.96	20.63	39.58
(15)	SylloBase	Avicenna	**79.71**	**69.80**	**77.42**
(16)	None	Avicenna	76.73	66.83	74.91

dation, and test sets with a ratio of 8:1:1 (*i.e.*, 800, 100, and 100 samples respectively). Then, we train models on the new training data and test their performance on the new test data. As a comparison, we also train models that have been pre-trained on the original training data (automatically constructed). The results are illustrated in Table 5.

It is clear to see that training on automatically constructed data is beneficial for learning manually rewritten data. This is due to the fact that the original dataset is large and contains sufficient training signals. This also validates the benefit of our dataset—the knowledge acquired from large-scale data can be transferred to more difficult problems.

Transfer Learning SylloBase supports study on five types of syllogisms. We explore their internal relationships through a transfer learning experiment. Besides, we also investigate if the knowledge learned on SylloBase can improve other syllogism datasets (*e.g.*, Avicenna). The results are shown in Table 6. In this experiment, we first train a BART model on one dataset (denoted as "pre-training"), then further train it on another dataset (denoted as "fine-tuning") and report the results.

In the first group of experiments (the first two rows), we can see learning categorical syllogisms contributes less to learning hypothetical and disjunctive syllogisms. This confirms our concern that merely studying categorical syllogisms is not enough, and it proves our contribution to syllogism study. In terms of the results in rows (3)-(9), we can generally conclude that learning basic syllogisms is beneficial for learning combined syllogisms, and vice versa. One exception is the result in the row (9), and it indicates that the knowledge learned from the complex syllogisms does not help for learning hypothetical syllogisms. We speculate the reasons are: (a) complex syllogisms have significantly more patterns than hypothetical syllogisms (42 vs. 3), and (b) the premise/conclusion of complex syllogisms is too complicated to form effective knowledge for hypothetical syllogisms. Finally, comparing the results in the row (15) and (16), we can see models trained on SylloBase have good generalizability on other syllogism datasets, demonstrating once again the value of our SylloBase on general syllogism research.

Table 7: Impact of context for conclusion generation (ROUGE-1/2/L).

Model	*w/o* Context	*w/* Context
GPT-2	34.38 / 11.07 / 30.42	22.33 / 5.16 / 19.44
T5	43.21 / 19.13 / 38.72	27.19 / 8.30 / 24.08
BART	41.85 / 18.20 / 37.59	25.71 / 8.02 / 22.71

Effect of Context in Premises Existing machine reading comprehension datasets often provide a paragraph for reasoning. Inspired by these tasks, we expand the premises in our generated syllogisms by adding more informative context so as to validate the models' capability of extracting effective clues and inferring conclusions. Specifically, for each premise in the manually rewritten dataset, we ask the annotators to further collect some relevant information through search engines and add it as the context. After this step, both premises are hidden in paragraphs, which makes it more difficult to infer a correct conclusion (as shown in Table 13). Results of both tasks shown in Table 7 indicate: (1) Existing models are still far from tackling reasoning problems in real life; and (2) Extracting clues (such as premises in our case) before reasoning is a promising solution for reasoning tasks, which could be explored in the future.

Appendix I shows a case study with some model-generated conclusions of syllogisms.

5 Conclusion

In this work, we built a large-scale benchmark for natural language syllogistic reasoning. It covers five types of syllogism. The data were automatically constructed from knowledge bases by our proposed construction methods. To evaluate the models' performance on real human syllogism, we manually rewrite 1,000 samples as the test set. Experiments show that syllogistic reasoning is a very challenging task for existing pre-trained language models. Moreover, our further study indicates that existing models are even farther from tackling syllogistic reasoning in real scenarios.

Ethical Statement

This work constructs a new benchmark for syllogistic reasoning. The main dataset is automatically constructed using entities and their relations from Wikidata and ConceptNet. The construction template is predefined and manually reviewed, so the ethical concerns are avoided. For the human rewriting process, we hire five annotators and require them to avoid any social bias and privacy issues in the rewritten material. The results are randomly shuffled and sent back to them for an ethical review. We pay them roughly $15 per hour for annotation.

Limitations

We build a new benchmark for syllogistic reasoning. The limitations are mainly in the experiments part: (1) Due to the limited human resources, our test set is quite small, which may not support training large models directly. (2) We evaluate all models by comparing their predictions with the ground-truth conclusions, but human performance is not evaluated. As a benchmark, it may be better to provide human performance and show the performance gap of existing models. (3) We have not tested the performance of pre-trained models in terms of logical correctness. This kind of automatic metrics has been rarely studied, which can be a potential direction of our future work.

References

Zeinab Aghahadi and Alireza Talebpour. 2022. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning. *Journal of Applied Non-Classical Logics*, 0(0):1–17.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 628–635. The Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Irving Copi, Carl Cohen, and Victor Rodych. 2016. *Introduction to logic*. Routledge.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Hannah Dames, Clemens Schiebel, and Marco Ragni. 2020. The role of feedback and post-error adaptations in reasoning. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*. cognitivesciencesociety.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail S. Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Li. 2022. Is GPT-3 a good data annotator? *CoRR*, abs/2212.10450.

Tiansi Dong, Chengjiang Li, Christian Bauckhage, Juanzi Li, Stefan Wrobel, and Armin B. Cremers. 2020. Learning syllogism with Euler neural-networks. *CoRR*, abs/2007.07320.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2391–2401. Association for Computational Linguistics.

William Huitt and John Hummel. 2003. Piaget's theory of cognitive development. *Educational psychology interactive*, 3(2).

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. 1990. CYC: toward programs with common sense. *Commun. ACM*, 33(8):30–49.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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 2020*, pages 3622–3628. ijcai.org.

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. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics, IWCS 2009, Tilburg, The Netherlands, January 7-9, 2009*, pages 140–156. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

Allen Newell and Herbert A. Simon. 1956. The logic theory machine-a complex information processing system. *IRE Trans. Inf. Theory*, 2(3):61–79.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2673–2679. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Shiya Peng, Lu Liu, Chang Liu, and Dong Yu. 2020. Exploring reasoning schemes: A dataset for syllogism figure identification. In *Chinese Lexical Semantics - 21st Workshop, CLSW 2020, Hong Kong, China, May 28-30, 2020, Revised Selected Papers*, volume 12278 of *Lecture Notes in Computer Science*, pages 445–451. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Adina Williams, Nikita Nangia, and Samuel R. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Algorithm 1: Human Paraphrasing Process

Input: an origin syllogism S, search engine
for premise or conclusion s *in* S **do**
  // Retrieval relevant sentences for premise or conclusion by search engine
  Retrieval(s) $\rightarrow$ retrieval result r;
  // Manually check if r can be used
  if ManualCheck(r) **then**
    $r \rightarrow s$;
  else
    ManualRewrite(r) $\rightarrow s$;
  end
end

Table 8: An example of paraphrasing process.

Original premise of a hypothetical syllogism **Premise**: Something that might happen as a consequence of attending a classical concert is going to sleep.
Retrieval and manual check **Premise**: I probably spend more concert time asleep than awake.
Rewriting **Premise**: When attending classical concerts, people probably spend more concert time asleep than awake.

A Patterns in Syllogism

We list all valid patterns in categorical (shown in Table 9), hypothetical (shown in Table 10), and complex syllogisms (shown in Table 11).

B Relations from Wikidata and ConceptNet

We list all relations that are used for constructing syllogisms in Table 12. For Wikidata, we use 16 relations, which are all used for constructing categorical syllogisms. As for ConceptNet, we use 15 relations, and they are used for constructing categorical, hypothetical, and disjunctive syllogisms.

C GPT-3 Rewriting

GPT-3 is a well-known pre-trained language model, which has demonstrated impressive few-shot performance on a wide range of natural language processing (NLP) tasks. Recently, researchers has tried to use GPT-3 to annotate data for NLP tasks (Ding et al., 2022). Inspired by this, we choose GPT-3 to complete the rewriting task. In our case, we use a prompt to ask GPT-3 to change the expression of the syllogism but keep its original meaning and pattern. We also append some human-rewritten examples in the prompt as few-shot input. The generated results have good quality in fluency, diversity, and logic, which are suitable for training models. The prompts used for rewriting are listed in Table 16-20.

D Human Rewriting

First, 500 samples are randomly collected from each type of syllogism, respectively. Then, we examine the semantics and filter out illogical syllogisms. Next, for the remaining ones, we correct the grammatical problems (if any). Finally, for each premise/conclusion, the language is painstakingly paraphrased. The paraphrasing process is illustrated in Algorithm 1, and an example is given in Table 8. After rewriting, the sample is more diverse, fluent, and closer to real human language.

E Annotation of Automatic Data

To evaluate the quality of our automatically generated data, we conduct a human annotation for 100 random samples (20 for each type of syllogisms). The annotators are asked to label whether the samples have grammatical faults and incorrect logic. The overall accuracy is 73%. Concretely, the accuracy is 70%, 90%, 70%, 65%, and 70% for categorical syllogisms, hypothetical syllogisms, disjunctive syllogisms, polysyllogisms, and complex syllogisms, respectively. This result reflects: (1) Our automatic data have fairly good quality. Our experiments in Section 4.4 also validates this. (2) The polysyllogism is hard to construct as it concerns multiple syllogisms.

F Distractor Construction in Conclusion Selection Task

In the conclusion selection task (introduced in Section 4.1), we mix the correct conclusion with three distractors. Basically, these distractors are generated from the ground-truth conclusion by changing its quantifier, adding negative words, or exchanging its subject and object. Specifically, for different kinds of syllogisms, we show the distractor generation process by some examples.

Categorical Syllogism For a syllogism as follows:

Premise 1: All m are p.
Premise 2: All s are m.
Conclusion: All s are p.

Table 9: 24 valid patterns in categorical syllogisms.

Pattern	Figure	Major premise	Minor premise	Conclusion
Barbara (AAA)	1	All m are p	All s are m	All s are p
Barbari (AAI*)	1	All m are p	All s are m	Some s are p
Celarent (EAE)	1	No m is p	All s are m	No s is p
Celaront (EAO*)	1	No m is p	All s are m	Some s are not p
Darii (AII)	1	All m are p	Some s are m	Some s are p
Ferio (EIO)	1	No m is p	Some s are m	Some s are not p
Camestres (AEE)	2	All p are m	No s is m	No s is p
Camestros (AEO*)	2	All p are m	No s is m	Some s are not p
Cesare (EAE)	2	No p is m	All s are m	No s is p
Cesaro (EAO*)	2	No p is m	All s are m	Some s are not p
Baroco (AOO)	2	All p are m	Some s are not m	Some s are not p
Festino (EIO)	2	No p is m	Some s are m	Some s are not p
Darapti (AAI)	3	All m are p	All m are s	Some s are p
Felapton (EAO)	3	No m is p	All m are s	Some s are not p
Datisi (AII)	3	All m are p	Some m are s	Some s are p
Disamis (IAI)	3	Some m are p	All m are s	Some s are p
Bocardo (OAO)	3	Some m are not p	All m are s	Some s are not p
Ferison (EIO)	3	No m is p	Some m are s	Some s are not p
Bamalip (AAI)	4	All p are m	All m are s	Some s are p
Calemes (AEE)	4	All p are m	No m is s	No s is p
Calemos (AEO*)	4	All p are m	No m is s	Some s ara not p
Fesapo (EAO)	4	No p is m	All m are s	Some s are not p
Dimatis (IAI)	4	Some p are m	All m are s	Some s are p
Fresison (EIO)	4	No p is m	Some m are s	Some s are not p

Table 10: Three valid patterns in hypothetical syllogism. P, Q, and R are three propositions.

Original hypothetical syllogism	
Premise 1: If P is true, then Q is true. Premise 2: If Q is true, then R is true.	Conclusion: If P is true, then R is true.
Modus ponens	
Premise 1: If P is true, then Q is true. Premise 2: P is true.	Conclusion: Q is true.
Modus tollens	
Premise 1: If P is true, then Q is true. Premise 2: Q is not true.	Conclusion: P is not true.

We can generate distractors of the conclusion as:

(1) Some s are p. (*modify quantifiers*)
(2) All s are not p. (*add negative words*)
(3) All p are s. (*exchange subjects and predicates*)
(4) Some p are not s. (*others*)

Hypothetical Syllogism For a syllogism as follows:

Premise 1: If P is true, then Q is true.
Premise 2: If Q is true, then R is true.
Conclusion: If P is true, then R is true.

We can generate distractors of the conclusion as:

(1) If R is true, then P is true. (*exchange propositions*)
(2) If Q is true, then P is true. (*exchange propositions*)
(3) If R is true, then Q is true. (*exchange propositions*)
(4) P is true. (*remove a proposition*)
(5) Q is true. (*remove a proposition*)
(6) R is true. (*remove a proposition*)
(7) If P is true, then R is not true. (*add negative words*)

Disjunctive Syllogism For a syllogism as follows:

Premise 1: P is true or Q is true;
Premise 2: P is not true;
Conclusion: Q is true.

Table 11: 42 valid patterns in complex syllogisms.

Id	Premise 1	Premise 2	Conclusion
0	$\neg p \vee q$	p	q
1	$(p \wedge q) \vee r$	$\neg p \vee \neg q$	r
2	$(p \vee q) \vee r$	$\neg p \wedge \neg q$	r
3	$p \vee \neg q$	$\neg p$	$\neg q$
4	$p \vee (q \wedge r)$	$\neg p \wedge q$	r
5	$p \vee (q \wedge r)$	$\neg p \wedge r$	q
6	$p \vee (q \vee r)$	$\neg p \wedge \neg r$	q
7	$\neg p \vee q$	$\neg q$	$\neg p$
8	$p \vee (q \vee r)$	$\neg q \wedge \neg r$	p
9	$(p \wedge q) \vee r$	$p \wedge \neg r$	q
10	$(p \wedge q) \vee r$	$q \wedge \neg r$	p
11	$p \vee \neg q$	q	p
12	$p \vee (q \wedge r)$	$\neg q \vee \neg r$	p
13	$\neg q \rightarrow \neg p$	$\neg q$	$\neg p$
14	$(p \vee q) \rightarrow r$	$p \vee q$	r
15	$(p \wedge q) \rightarrow r$	$p \wedge q$	r
16	$p \rightarrow (q \vee r)$	p	$q \vee r$
17	$p \rightarrow (q \vee r)$	$p \wedge \neg q$	r
18	$p \rightarrow (q \vee r)$	$p \wedge \neg r$	q
19	$p \rightarrow (q \wedge r)$	p	$q \wedge r$
20	$p \rightarrow (q \wedge r)$	$p \wedge q$	r
21	$p \rightarrow (q \wedge r)$	$p \wedge r$	q
22	$(p \vee q) \rightarrow r$	$\neg r$	$\neg (p \vee q)$
23	$(p \vee q) \rightarrow r$	$\neg p \wedge \neg r$	$\neg q$
24	$(p \vee q) \rightarrow r$	$\neg q \wedge \neg r$	$\neg p$
25	$(p \wedge q) \rightarrow r$	$\neg r$	$\neg (p \wedge q)$
26	$(p \wedge q) \rightarrow r$	$p \wedge \neg r$	$\neg q$
27	$(p \wedge q) \rightarrow r$	$q \wedge \neg r$	$\neg p$
28	$p \rightarrow (q \vee r)$	$\neg q \wedge \neg r$	$\neg p$
29	$p \rightarrow (q \wedge r)$	$\neg q \vee \neg r$	$\neg p$
30	$\neg q \rightarrow \neg p$	$\neg r \rightarrow \neg q$	$\neg r \rightarrow \neg p$
31	$(p \vee q) \rightarrow r$	$r \rightarrow s$	$(p \vee q) \rightarrow s$
32	$(p \vee q) \rightarrow r$	$(r \rightarrow s) \wedge p$	s
33	$(p \vee q) \rightarrow r$	$(r \rightarrow s) \wedge q$	s
34	$(p \wedge q) \rightarrow r$	$r \rightarrow s$	$(p \wedge q) \rightarrow s$
35	$(p \wedge q) \rightarrow r$	$(r \rightarrow s) \wedge p \wedge q$	s
36	$p \rightarrow (q \vee r)$	$(q \vee r) \rightarrow s$	$p \rightarrow s$
37	$p \rightarrow (q \wedge r)$	$(q \wedge r) \rightarrow s$	$p \rightarrow s$
38	$p \rightarrow q$	$q \rightarrow (r \vee s)$	$p \rightarrow (r \vee s)$
39	$p \rightarrow q$	$(q \rightarrow (r \vee s)) \wedge p$	$r \vee s$
40	$p \rightarrow q$	$q \rightarrow (r \wedge s)$	$p \rightarrow (r \wedge s)$
41	$p \rightarrow q$	$(q \rightarrow (r \wedge s)) \wedge p$	$r \wedge s$

Table 12: Relations used for syllogisms construction.

Type	Used Relations
Wikidata	
Categorical	academic degree subclass (human)
Categorical	ethnic subclass (human)
Categorical	field of work subclass (human)
Categorical	genre subclass (human)
Categorical	occupation subclass (human)
Categorical	language subclass (human)
Categorical	instance of (human)
Categorical	instance of (taxon)
Categorical	taxon subclass (taxon)
Categorical	film subclass (film)
Categorical	chemical compound subclass (chemical compound)
Categorical	administrative territorial subclass (administrative territorial)
Categorical	architectural structure subclass (architectural structure)
Categorical	astronomical object subclass (astronomical object)
Categorical	occurrence subclass (occurrence)
Categorical	thoroughfare subclass (thoroughfare)
ConceptNet	
Categorical / Disjunctive	/r/CapableOf
Categorical / Disjunctive	/r/HasProperty
Categorical / Disjunctive	/r/Antonym
Categorical / Disjunctive	/r/DistinctFrom
Disjunctive	/r/Part of
Disjunctive	/r/HasA
Disjunctive	/r/UsedFor
Disjunctive	/r/SymbolOf
Disjunctive	/r/MannerOf
Disjunctive	/r/MadeOf
Hypothetical	/r/Causes
Hypothetical	/r/HasSubevent
Hypothetical	/r/HasPrerequisite
Hypothetical	/r/MotivatedByGoal
Hypothetical	/r/CausesDesire

We can generate distractors of the conclusion as:

(1) Q is not true. *(add negative words)*
(2) P is true. *(change a proposition)*
(3) P is true or Q is not true. *(add a proposition)*

Polysyllogism Syllogism This kind of syllogism is built on several categorical syllogisms. Therefore, we can use the same distractor construction method as categorical syllogisms.

Complex Syllogism This kind of syllogism is constructed by adding one or model logical connectives to the original premises and conclusions. Therefore, to generate the distractors, we can (1) add or remove the negative connective (*i.e.*, *not*) from the original proposition; or (2) replace the connectives in the original proposition by others (*e.g.*, *and* $\rightarrow$ *or*). For example, given a syllogism as follows:

Premise 1: If P is true or if Q is true, then R is true;
Premise 2: If R is true, then S is true;
Conclusion: If P is true or if Q is true, then S is true.

Table 13: An example of syllogism with context. The vanilla premises are in orange.

Premise 1: Carbon dioxide is a chemical compound composed of two oxygen atoms covalently bonded to a single carbon atom. CO2 exists in the earth's atmosphere as a gas and in its solid state it known as dry ice. **Premise 2:** In a scientific context, "pure" denotes a single type of material. Ostensibly, compounds contain more than one type of material. Therefore, chemical compounds are considered pure substances. Pure compounds are created when elements combine permanently, forming one substance. **Conclusion:** Pure substances include carbon dioxide.

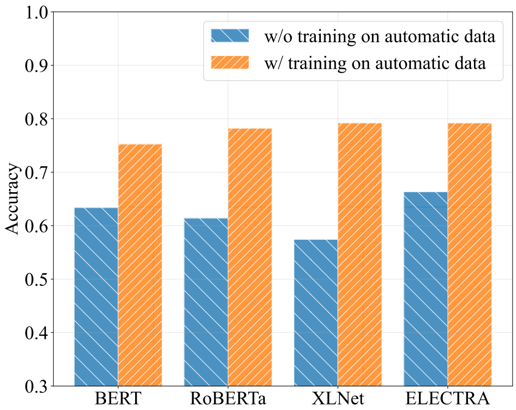

Figure 2: Results of the conclusion selection task with or without pre-training on automatic training data.

We can generate distractors of the conclusion as:

(1) If P is true or if Q is true, then S is not true. *(add negative words)*

(2) If P is true or if S is true, then Q is true. *(change a proposition)*

(3) If P is true and if S is true, then Q is true. *(change the logical connective words)*

G Dataset Statistics

The statistics of our SYLLOBASE is given in Table 14.

H Implementation Details

We use PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2019) to implement all models. They are trained on 8 Tesla V100 GPUs with 32GB memory. All hyperparameters (*e.g.*, learning rate) are tuned according to the performance (BLEU-1/Accuracy) on the validation set.

In the conclusion generation task, for the decoder-only model GPT-2, the major premise and minor premise are concatenated as a long sequence and fed into the model (decoder) to generate the conclusion. For the encoder-decoder structure (Transformer, T5, and BART), the two premises are concatenated and input to the encoder, while the conclusion is input to the decoder and used for generation. The maximum generation length is set as 128. The training batch size is set as 32. The AdamW (Loshchilov and Hutter, 2019) optimizer is applied with a learning rate of 5e-5. The learning rate decay mechanism is applied. All models are trained by 10 epochs, and the total training time is around 1.22 hours.

In the conclusion selection task, we concatenate two premises as one sequence, use the conclusion as another sequence, and transform them into the text-pair input format, which is commonly supported by pre-trained language models. For example, the input for BERT is: $X =$ `[CLS]`P_1P_2`[SEP]`C`[SEP]`. The representation of `[CLS]` is used for option selection. The maximum sequence length is set as 256. The training batch size is set as 64. A learning rate of 2e-5 with decay mechanism is used. The optimizer is also AdamW. All models are trained by ten epochs, and the total training time is around 3.29 hours.

I Case Study

We show some results of BART in conclusion generation task to make a case study. We have listed a good case and a bad case for each type of syllogism. They are shown in Table 21. We can see: (1) The model can generate conclusions that are different from the ground-truth but are also correct in logic. This indicates that pre-trained language models can indeed learn some logic reasoning skills from syllogisms rather than merely "remembering" some fixed patterns. (2) Syllogistic reasoning is still difficult for existing models, and the errors stem from several different aspects. As shown in the hypothetical syllogism, the model generates a semantically correct conclusion, but it is irrelevant to the premises. This problem is identified as "hallucination" of pre-trained language models (Nie et al., 2019), *i.e.*, the model cannot decide whether to generate a conclusion based on its learned parameters or the given context. We believe our dataset can contribute to the study of hallucinations in logical reasoning. As for the last case, the model generates a conclusion opposite to the ground-truth. This indicates that existing models may need additional reasoning modules to conduct complex reasoning problems.

Table 14: Statistics of SYLLOBASE.

Conclusion Generation	Training	Validation	Test (*w/o* context)	Test (*w/* context)
# Premises-Conclusion Pair	40,000	10,000	1,000	1,000
Avg./Max. # Tokens in Premises	33.73 / 115	33.83 / 105	27.59 / 75	183.92 / 726
Avg./Max. # Tokens in Conclusion	11.84 / 66	11.91 / 62	8.5 / 21	8.5 / 21
Conclusion Selection	**Training**	**Validation**	**Test (*w/o* context)**	**Test (*w/* context)**
# Premises-Question Pair	40,000	10,000	1,000	1,000
Avg./Max. # Tokens in Premises	33.73 / 115	33.83 / 105	27.59 / 75	183.92 / 726
Avg./Max. # Tokens in Question	12.39 / 16	12.39 / 16	12.38 / 16	12.38 / 16
Avg./Max. # Tokens in Candidate Answer	11.53 / 71	11.50 / 64	9.41 / 26	9.41 / 26

Table 15: Results of conclusion generation task on validation set. "R-1/2/L" stands for Rouge-1/2/L, "B-1/2" stands for BLEU-1/2, and "BS" denotes BERT-Score.

	R-1	R-2	R-L	B-1	B-2	BS	R-1	R-2	R-L	B-1	B-2	BS	R-1	R-2	R-L	B-1	B-2	BS
Model	Categorical						Hypothetical						Disjunctive					
Transformer	16.85	3.63	14.95	7.38	1.4	83.09	24.02	5.67	22.29	19.36	4.54	87.32	16.74	2.72	15.47	8.99	1.34	83.95
GPT-2	30.36	8.68	27.41	26.87	7.55	89.05	31.51	9.8	28.96	26.06	7.61	90.68	32.11	9.86	29.53	24.91	7.21	90.34
T5	34.63	12.12	31.53	**31.65**	**11.21**	89.68	**36.99**	**14.92**	**34.69**	**32.7**	13.04	91.52	**40.75**	**18.2**	**38.36**	**35.05**	**16.24**	**91.87**
BART	**35**	**12.27**	**31.69**	30.76	10.7	**89.78**	36.44	14.84	34.32	32.33	**13.09**	**91.56**	40.62	18.13	38.26	34.67	15.83	91.79
Model	Polysyllogism						Complex						All					
Transformer	31.23	10.28	29.16	17.77	5.15	87.36	20.36	4.71	19.11	10.14	1.99	85.61	24.59	6.25	22.65	18.68	4.48	87.22
GPT-2	49.23	24.32	46.22	41.37	19.3	91.87	36.25	14.18	33.83	27.87	9.39	90.72	36.16	13.63	33.52	29.4	10.1	90.59
T5	55.79	30.72	53.01	**51.99**	28.74	92.98	43.93	22.84	41.81	38.01	19.13	91.93	**42.86**	19.99	**40.23**	**37.94**	**17.51**	91.69
BART	**56.49**	**31.23**	**53.52**	51.78	**28.79**	**93.14**	**45.21**	**23.99**	**42.81**	**38.35**	**19.6**	**92.14**	42.85	**20.26**	40.17	37.3	17.44	**91.75**

Table 16: GPT-3 rewriting prompts for categorical syllogisms.

Rewrite the following sentences to standard English. Keep the meaning and pattern of the original sentences, but change the expression of the sentences.

pattern: All m are p. Some s are m. [Therefore], some s are p.
original sentences: All sugar are carbohydrate. Some decay teeth are sugar. [Therefore], some decay teeth are carbohydrate.
rewritten sentences: Sugars are carbohydrates. Somethings that decay your teeth are sugary foods and drinks. [Therefore], carbohydrate eating can sometimes promote tooth decay.
pattern: Some p are m. All m are s. [Therefore], some s are p.
original sentences: Some visual art are art of painting. All art of painting are activity. [Therefore], some activity are visual art.
rewritten sentences: The visual arts are art forms that create works that are primarily visual in nature, such as painting. Painting is the practice of applying paint, pigment, color or other medium to a solid surface. [Therefore], creativily activties are used to develop new artistic works, such as visual art.
pattern: No p is m. All m are s. [Therefore], some s are not p.
original sentences: No animal is mineral. All mineral are solid. [Therefore], some solid are not animal.
rewritten sentences: Evidently, animal and vegetables are living, minerals not. A mineral is a naturally occurring inorganic solid. [Therefore], some substances are solids and are not living beings.
pattern: No p is m. All s are m. [Therefore], No s is p.
original sentences: No animal is plant. All rose are plant. [Therefore], no rose is animal.
rewritten sentences: Traditionally, Animals cannot produce their own energy which not like plants. A rose is a woody perennial flowering plant of the genus Rosa. [Therefore], roses and animals are extremely different species in nature.
pattern: No m is p. All s are m. [Therefore], some s are not p.
original sentences: No art is clumsiness. All sculpture are art. [Therefore], Some sculpture are not clumsiness.
rewritten sentences: Clumsiness is the lack of gracefulness or skill, whereas Art encompasses a diverse range of skill and techniques. Sculptures are artworks crafted with various media and materials. [Therefore], Sculpture makers are often talented at expressing creativity and not clumsy.

Table 17: GPT-3 rewriting prompts for hypothetical syllogisms.

Rewrite the following sentences to standard English. Keep the meaning and pattern of the original sentences, but change the expression of the sentences. **pattern:** If P is true, then Q is true. If Q is true, then R is true. [Therefore], if P is true, then R is true. **original sentences:** Something you might do while dating is kiss. Something that might happen when you kiss someone is they smile. [Therefore], something that might happen when you dating is they smile. **rewritten sentences:** When you are dating your beloved, you might have a sweet kiss. When you kiss your partner, you may find yourself smiling. [Therefore], when you are dating, you may find that you always have a smile on your face. **pattern:** If P is true, then Q is true. If Q is true, then R is true. [Therefore], if P is true, then R is true. **original sentences:** The effect of diminishing your own hunger is eating. The effect of eating is a full stomach. [Therefore], the effect of diminishing your own hunger is a full stomach. **rewritten sentences:** We are all aware that in order to reduce our hunger, we must consume food. Having a belly stuffed with comforting food can feel like a warm hug from the inside. [Therefore], We may feel full after we have satisfied our appetite. **pattern:** If P is true, then Q is true. If Q is true, then R is true. [Therefore], if P is true, then R is true. **original sentences:** Because you want to enjoy yourself, you would listen to music. Because you want to listen to music, you would hear singing. [Therefore], because you want to enjoy yourself, you would hear singing. **rewritten sentences:** If you want to enjoy yourself after a long day of work, you may listen to music. You want to hear your favorite musician sing because you appreciate music. [Therefore], because you want to have fun, you want to hear some singing. **pattern:** If P is true, then Q is true. If Q is true, then R is true. [Therefore], if P is true, then R is true. **original sentences:** attending a lecture requires you to listen. If you want to listen then you should not talk so much yourself. [Therefore], If you want to attending a lecture then you should not talk so much yourself. **rewritten sentences:** If you are in a lecture, you should focus your attention on listening and be mindful of disrupting the session due to speaking. If you want to devote your time to following the lecture without distractions, you must be aware of reducing your own babble. [Therefore], If you desire to remain in the lecture, it is key to dial down your chatter.

Table 18: GPT-3 rewriting prompts for disjunctive syllogisms.

Rewrite the following sentences to standard English. Keep the meaning and pattern of the original sentences, but change the expression of the sentences. **pattern:** P is true or Q is true. P is not true. [Therefore], Q is true. **original sentences:** Is the meal hot or cool. The meal are not hot. [Therefore], the meal are cool. **rewritten sentences:** The meal is warm or cold when the man gets home from work. The food is not warm when the man stays late at work. [Therefore], the meal is cold when the man comes home late. **pattern:** P is true or Q is true. P is not true. [Therefore], Q is true. **original sentences:** The ocean is gas or liquid. The ocean is not gas. [Therefore], the ocean is liquid. **rewritten sentences:** The ocean can exist in either liquid or gaseous form. The ocean is not gaseous. [Therefore], oceans do not exist in a gaseous condition, as far as we know. **pattern:** P is true or Q is true. P is not true. [Therefore], Q is true. **original sentences:** Memories are good or sad. Memories are not good. [Therefore], memories are sad. **rewritten sentences:** People like being engrossed in memories, whether good or sad. Old memories are not always pleasant. [Therefore], memories of the past may cause sadness. **pattern:** P is true or Q is true. P is not true. [Therefore], Q is true. **original sentences:** You can use an audience to performing in front of or boost your ego. You can not use an audience to boost your ego. [Therefore], you can use an audience to performing in front of. **rewritten sentences:** When you're in front of an audience, you can put on a show or increase your self-esteem. You cannot exaggerate your ego in front of an audience. [Therefore], you can give a performance in front of an audience. **pattern:** P is true or Q is true. P is not true, [Therefore], Q is true. **original sentences:** My flowers are ugly or pretty. My flowers are not ugly. [Therefore], My flowers are pretty. **rewritten sentences:** The blooms in my garden are either comely or unappealing. The blooms in my garden are not unsightly. Therefore, These flowers are indeed attractive.

Table 19: GPT-3 rewriting prompts for polysyllogisms.

Rewrite the following sentences to standard English. Keep the meaning of the original sentences, but change the expression of the sentences.

original sentences: No hypothesis is fact. Some proposition are hypothesis. Some proposition are not fact. All proposition are abstract object. [Therefore], some abstract object are not fact.
rewritten sentences: A hypothesis is a proposed explanation that differs from fact. Some propositions are hypotheses. Some propositions are proven not to be facts. Every proposition is an abstract object. [Therefore], some abstract objects do not exist as facts.
original sentences: Applied science is science. No Science is art. Human science is science. Some Behavioral genetics are not human science. Behaviour genetics is psychology. Genetics is biology. [Therefore], some applied science are not biology.
rewritten sentences: Applied science is science in every sense of the word. Science and art are two distinct forms of scholarship. Human science is a branch of science. Behavioral genetics does not involve any human science. Behavioral genetics is a branch of psychology. Genetics is the study of biology. [Therefore], applied science encompasses more than just biology.
original sentences: All feline are animal. no plant are animal. All flowering plants are plants. All tiger are genus Panthera. [Therefore], no Panthera are flowering plants.
rewritten sentences: A feline is an animal belonging to the cat family. There are many obvious differences between plants and animals. Flowering plants are plants that produce flowers and fruits. The tiger is a member of the genus Panthera. [Therefore], Panthera is different from flowering plants.
original sentences: All medication are drug. All hormone are medication. All plant hormone are hormone. Some plant hormone are gibberellins. All drug are useful. All gibberellins are carboxylic acid. [Therefore], Some carboxylic acid are useful.
rewritten sentences: A medication is a type of drug. Hormones are a type of medication. Plant hormones are a subset of hormones. Gibberellins are one type of plant hormone. All drugs have some sort of usefulness. Gibberellins are carboxylic acids. [Therefore], some carboxylic acids can be useful.

Table 20: GPT-3 rewriting prompts for complex syllogisms.

Rewrite the following sentences to standard English. Keep the meaning of the original sentences, but change the expression of the sentences.

original sentences: If you want to eat then you should open the refrigerator and open the chiller. It is eat, and it is open the refrigerator. [Therefore], it is open the chiller.
rewritten sentences: If you want to eat something after working out, open the refrigerator and make sure the chiller is in good working order. When you are hungry and want to eat anything, you may open the refrigerator to find some food. [Therefore], you must ensure that the chiller is operational.
original sentences: My toes are warmth and affectionateness, or My toes are cold. My toes are not warmth or not affectionateness. [Therefore], my toes are cold.
rewritten sentences: The temperature here varies greatly, my toes can be warm and friendly, or they might be freezing. Because of the low temperature, my toes are not warm or loving. [Therefore], my toes are freezing.
original sentences: Neon lights are dark, or Neon lights are bright and burnished. Neon lights are not dark, and Neon lights are burnished. [Therefore], Neon lights are bright.
rewritten sentences: Neon lights might be dim at times, but most of the time they are dazzling and burnished at night. Neon lights are not dim, and they shine burnished at night. [Therefore], the neon lights are really bright at night.
original sentences: Something you might do while not getting exercise is not get tired. Sometimes you don't getting exercise. [Therefore], You don't get tired.
rewritten sentences: If you do not exercise, you might remain energetic. When you don't workout occasionally, you will not become exhausted. [Therefore], If you are not exercising you will not get tired.

Table 21: Examples of conclusions generated by BART. The good results and bad results are in different colors.

Categorical Syllogism **Premise 1:** Carbon dioxide is a chemical compound. **Premise 2:** Chemical compounds are considered pure substances. **Conclusion:** Pure substances include carbon dioxide. **Prediction**: Some completely natural substances include carbon dioxide.
Categorical Syllogism **Premise 1:** All machines are not human. **Premise 2:** Every truck crane is a type of apparatus. **Conclusion:** All truck cranes are not human. **Prediction**: Some apparatus are not truck cranes.
Hypothetical Syllogism **Premise 1:** What you might do in class is sleep. **Premise 2:** A dream is something that can happen to you while you retire for the night. **Conclusion:** You're most likely daydreaming in class. **Prediction**: When you attend class, you may have a dream.
Hypothetical Syllogism **Premise 1:** When you make progress in your project, you may want to celebrate. **Premise 2:** Having a party is a good choice if you want to celebrate. **Conclusion:** You may want to have a party if you achieve great progress in your project. **Prediction**: Celebrating has the effect of bringing people together.
Disjunctive Syllogism **Premise 1:** Newspapers are generally published daily or weekly. **Premise 2:** Some newspapers are not published weekly. **Conclusion:** Some newspapers are daily newspapers. **Prediction**: Some newspapers are generally published daily.
Disjunctive Syllogism **Premise 1:** There is still controversy about whether there are ghosts in the world. **Premise 2:** The man is an atheist. **Conclusion:** The man thinks that there is no ghost in the world. **Prediction**: There are ghosts in the world.
Polysyllogism **Premise 1:** Some movies are not cartoon movies. **Premise 2:** Science fiction animations belong to animated films. **Premise 3:** Remake films are also films. **Conclusion:** Some remakes are out of scope of science fiction cartoons. **Prediction**: It exists some remakes are out of scope of science fiction cartoon.
Polysyllogism **Premise 1:** All Pierce Brosnan are human. **Premise 2:** No human is a beast. **Premise 3:** Some Pierce Brosnan are not machines. **Conclusion:** Some machines are not beasts. **Prediction**: Some Pierce Brosnan are not beasts.
Complex Syllogism **Premise 1:** Fencing requires wearing a protective mask and gloves. **Premise 2:** The woman is fencing and wearing a wiremesh mask. **Conclusion:** The woman is also wearing gloves. **Prediction**: The woman may be wearing gloves.
Complex Syllogism **Premise 1:** If Jack has computer skills *and* programming knowledge, he could write programs. **Premise 2:** Jack cannot write computer programs, but he can use computers. **Conclusion:** Jack does not have programming knowledge. **Prediction**: He can write computer programs.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitation (Page 9).

☑ A2. Did you discuss any potential risks of your work?
We have an ethical statement section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction (Page 1).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Footnotes in Section 2.2 and Section 3.1 and the References section.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No, because the licenses are well known, which allow use of the artifacts in work like ours.

☒ 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)?
No, the use of existing artifacts is only for research, not for commercial use.

☑ 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?
Section 3 Data Construction (Page 3).

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We note the language and domains in section 3.1 Data Source (Page 3).

☑ 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.
Section 3 Data Construction (Page 3) and Appendix G Data Statistics (Page 15).

C ☑ Did you run computational experiments?

Appendix H Implementation Details (Page 15).

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix H Implementation Details (Page 15).

☑ 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?
Section 4.3 Experimental Results.

☒ 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.)?
We used the transformers package for training baseline models and the ROUGE, BLEU, BERT-Score packages for evaluation. Since we only used the common functions/interfaces well known in the NLP, we did not discuss the details.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Appendix D Human Rewriting.

☒ 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, the human annotators are required to just rewrite the automatic samples, which is unnecessary to give a instruction.

☑ 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)?
Section Ethical Statement.

☒ 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, the human annotators are required to just rewrite the automatic data, which is unnecessary to give a instruction.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our benchmark has no social impacts and we just use some open knowledge bases, like ConceptNet and Wikidata. There is no need to get the approval of an ethics review board.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No, we required the annotators to avoid any social bias and privacy issues in the rewritten material, which is discussed in the Section Ethical Statement.