ILogicEval: A Counterintuitive Logical Reasoning Evaluation Dataset

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have garnered significant attention worldwide due to their increasing size and improving capabilities. However, as LLMs continue to expand, traditional benchmark datasets are becoming less effective in evaluating their reasoning skills. This is primarily due to the difficulty of the tasks and issues with data contamination. Meanwhile, in the domain of logical reasoning, existing benchmarks often lack the ability to isolate specific reasoning abilities and fail to provide sufficient evidence for answer derivation. To address these issues, a novel dataset ILogicEval is proposed, which consists of sentences composed of unrelated statements, challenging LLMs to answer questions that cannot be solved based on their learned knowledge. ILogicEval is carefully designed to incorporate rich language diversity and assess the logical reasoning ability of LLMs independently of other reasoning skills, such as commonsense reasoning. To ensure a more reliable evaluation, we also introduce a new evaluation metric that mitigates the influence of bias and randomness inherent in LLMs. Through experiments, we demonstrate the extent to which logical reasoning is required to answer the questions in ILogicEval and compare the performance of different popular LLMs in conducting logical reasoning. This dataset and evaluation metric address the limitations of existing benchmarks, providing a comprehensive assessment of the logical reasoning capabilities of LLMs.

1 Introduction

Large Language Models (LLM) have been evaluated on various datasets to assess different abilities, including mathematics reasoning, instruction following, and code generation (Cobbe et al., 2021; Wang et al., 2023; Chen et al., 2021). Logic reasoning, has long been utilized as an important measure for evaluating human intelligence level during onboarding, graduate admission, and civil servant recruitment. Many globally recognized examinations usually include a significant portion of logic puzzles, such as the LSAT, GMAT, civil service examinations, and aptitude tests.

Despite numerous datasets that exist for evaluating logical reasoning, they are no longer effective for assessing the current LLMs due to their current capabilities and data leakage during pretraining. TS-Guessing (Deng et al., 2023) estimates data leakage severity by predicting critical keywords and incorrect option predictions on testing sets, revealing probable data contamination in LLM evaluation benchmarks. Notably, the authors highlight the dataset TruthfulQA, which was released subsequent to GPT-3.5-turbo, yet still achieves an accuracy of over 50% in guessing incorrect options. Evidence of data contamination across different benchmarks has also been found and recorded in the LM contamination index. Given the vast amount of pretraining data, preventing data contamination is unrealistic. Thus, we propose a novel dataset, ILogicEval, constructed using contaminated datasets in a counterintuitive manner to address this issue. For example, while GPT4 achieves over 80% accuracy on the original datasets used to create ILogicEval (SNLI, MNLI, and ReClor), its performance drops to only 32.2% on ILogicEval.

Furthermore, as LLMs continue to be trained on ever-expanding datasets, they may improve their accuracy on established reasoning tests without necessarily enhancing their reasoning skills (Tian et al., 2023). This poses a challenge to the ongoing evaluation of specific reasoning abilities over time using conventional benchmark datasets. ILogicEval tackles this challenge in the logic reasoning domain by providing content that LLMs are unlikely to encounter during pretraining. For example, it includes statements connected in ways that rarely occur in daily conversations. By relying on unfamiliar concepts and relationships, ILogicEval aims to minimize the potential influence of data

contamination when evaluating logical reasoning over time.

Some studies (Yu et al., 2020; Liu et al., 2020) focus on general evaluation in the domain of logical reasoning, where the chain for explaining the answer correctness is more complex and challenging to retrieve. These evaluation benchmarks also typically do not concentrate on a single domain reasoning ability, various reasoning skills can contribute to solving the task,making it challenging to pinpoint the specific abilities of LLMs. On the other hand, some works (Clark et al., 2020; Sanyal et al., 2022; Han et al., 2022) primarily focus on assessing a single reasoning ability while minimizing the contribution of other abilities, such as commonsense reasoning. However, these works have received less attention from the research community, possibly due to their limited language diversity, which hinders the reflection of model performance in more general scenarios. To address this limitation, we propose a new evaluation benchmark "**I**llogical" **Logic**al Reasoning **Eval**uation (ILogicEval), a logical reasoning-focused synthetic dataset with rich language diversity for evaluating LLMs.

In summary, this paper makes several key contributions: (1) It reduces the possibility of data contamination in assessing the logical reasoning capabilities of LLMs over time by constructing counterintuitive sentences that deviate from real-life scenarios. (2) It minimizes the involvement of other reasoning abilities by formulating the dataset as symbolic logical propositions before translating them into counterintuitive sentences in natural language. (3) It enhances language diversity by sampling from corpora with extensive vocabularies, enabling a more challenging and comprehensive evaluation. (4) It proposes a new evaluation metric that considers the confidence level of models, facilitating a more comprehensive assessment of the logical reasoning ability of LLMs.

2 Related Work

2.1 Traditional Machine Reading Comprehension Benchmarks

In the field of evaluating LLMs, numerous popular benchmarks have been constructed, each designed to assess different reasoning capabilities. For instance, TriviaQA (Joshi et al., 2017) examines the ability of models to reason across multiple documents. HotpotQA (Yang et al., 2018) evaluates the multi-hop reasoning ability of models. Drop (Dua et al., 2019) assesses their discrete numerical reasoning ability and MuTual (Cui et al., 2020) evaluates their ability to reason with dialogues. These datasets are commonly sourced from online content and involve various question types ranging from span retrieval tasks to more complex reasoning tasks. However, LLMs may have been exposed to similar content or question types during pre-training, allowing them to achieve high accuracy without genuinely improving their underlying reasoning abilities (Tian et al., 2023). To address this issue, we propose ILogicEval which is counterintuitive to everyday contexts. Similar content is unlikely to exist online and to be learned by LLMs. This allows ILogicEval to provide a more effective evaluation on the logical reasoning capabilities of LLMs.

2.2 Complex Logical Reasoning Datasets

Various datasets have been introduced to evaluate the reasoning ability of LLMs at a more domain-specific level. In the logic reasoning domain, there are two notable multiple-choice question answering (MCQA) datasets that are composed of inference questions. ReClor (Yu et al., 2020) is derived from GMAT and LSAT questions while LogiQA (Liu et al., 2020) is sourced from the Chinese Civil Servants Examination. These datasets concern more than just inference problems, the correct derivation of answers may involve commonsense reasoning, allowing LLMs to leverage their inherent knowledge learned during pre-training to answer the questions. It remains unclear whether a performance increase can be attributed to the enhanced ability in commonsense reasoning or logical reasoning. ILogicEval addresses this issue with its counterintuitive content and content construction grounded in propositional logic, making commonsense knowledge likely to be inapplicable in answer generation.

2.3 First-Order Logic Reasoning Datasets

There are numerous domain-specific benchmarks that focus on inference problems. For instance, Ruletaker (Clark et al., 2020), LogicNLI (Tian et al., 2021) and RobustLR (Sanyal et al., 2022) are synthetic datasets created to assess various aspects of model performance, such as accuracy, robustness, generalization, and traceability.On the other hand, FOLIO (Han et al., 2022) is a dataset constructed under human supervision, which aims to establish a dataset with complex logical reasoning

structure and a richer vocabulary compared to other first-order logic reasoning datasets. However, its distinct vocabulary size remains in the thousands, which is incomparable to the complex logical reasoning datasets. To address this issue, ILogicEval derived from NLI and SNLI data that presents more complex language structures with a vocabulary size in the tens of thousands level, provides a more challenging and comprehensive evaluation of a model's capabilities.

3 Dataset Construction and Overview

3.1 Dataset Construction

To construct a dataset with supportive explanations, ILogicEval is constructed based on verifiable propositional logic. The dataset was initially created using a set of randomly generated logical propositions, and subsequently converted into natural language subsequently to form the MCQA dataset. Previous studies have investigated the models' ability to solve logical inference puzzles in a symbolic form (Hahn et al., 2021; Pi et al., 2022) and other works have converted symbolic propositions into natural language using a predefined set of subjects and adjectives (Clark et al., 2020; Tian et al., 2021; Sanyal et al., 2022). This approach ensures that the generated natural language datasets adhere to the rules of logical inference.

In contrast to the previous approaches, our method involves the random construction of expressions that incorporate noise information, which does not contribute to the inference derivation. Additionally, we sample sentences in high language diversity and link generally unrelated sentences despite their usual lack of co-occurrence. This approach ensures the effectiveness of evaluation over time while introducing a certain level of difficulty in logical reasoning.

To facilitate evaluation and enhance user experience, we propose a multiple-choice MRC dataset consisting of three components: content, passage, and four options, with one option being correct. The four options are generated by an external validator, which examines the entailment between the content and each option. The multiple-choice approach ensures ease and effectiveness in evaluation.

3.1.1 Formation of Symbolic Logical Propositions

The content part of the MRC dataset is composed of multiple logical propositions. Each proposition is derived from a random selection of three logical variables from a set of eight possible variables (i.e., 'A', 'B', ..., 'H'). The probability of selection is computed along the content formation process as illustrated in Algorithm 1 in Appendix. The value of n is set to 3. After variables selection, they are incorporated into the premise of the following three implication rules commonly used in previous research to address (Wang et al., 2022; Li et al., 2022; Zhao et al., 2022) or generate (Clark et al., 2020; Sanyal et al., 2022) logical reasoning benchmarks.

$$((A \to B) \to (\neg B \to \neg A)) \quad (1a)$$

$$((\neg(A \wedge B) \to C) \to (\neg A \to C)) \quad (1b)$$

$$(((A \vee B) \to C) \to (A \to C)) \quad (1c)$$

In case the third rule is chosen, only the first two variables are utilized and the last variable is discarded.

In order to avoid sentence repetition and prevent the inference of answers from multiple constructed propositions, a limitation is imposed on the maximum occurrence of a logic variable within a single instance. With the maximum number of propositions of n, if a variable appears more than $n-1$ times in the content, its probability of selection is set to 0.1. If the variable is selected n times or more, its probability is set to zero. Otherwise, the probability of the i-th variable is calculated as follows,

$$\frac{\max(o) + 1 - o_i}{\sum_i (\max(o) + 1 - o_i)} \quad (2)$$

where o denotes an array that contains the occurrence of all eight variables in the constructed propositions used for constructing the content of a single instance and $i \in (0, 8)$. To construct an instance, two to $n+1$ logical propositions are generated to form the content part of the MCQA dataset.

After constructing the content for MCQA, the generated propositions q and the variable picking counter information o are utilized to construct the option sets of MCQA. The set of variables x' is retrieved at first, where $0 < o_k < n$ for $k \in x'$. This retrieval ensures that the variables in x' have been selected in lower occurrences, thus increasing the difficulty of the questions. With the variables in x', the options set is created, which includes all four negation versions of all possible pairs of variables (i.e., $(A \to B)$, $(\neg A \to B)$, $(A \to \neg B)$, and $(\neg A \to \neg B)$), as well as two negation versions of a single variable in x' (i.e., $\neg A$ and A).

The set is then divided into the entailment group e and the non-entailment group n using an external logic validator. The entailment group contains the conclusions entailed by the content propositions q, while the non-entailment group contains the conclusions that are not entailed. To ensure that answers cannot be inferred from a single proposition, the answers of variable pairs in the entailment group e that can be directly derived from a single proposition q_i, where $q_i \in q$ in the content part, are filtered out. This filtering ensures that at least two propositions in the content are necessary to derive the correct answer.

3.1.2 Formation of Question Types

Critical thinking assessment is widely acknowledged as a crucial skill for reasoning, and be utilized as an indicator in prestigious examinations for both career and academic progression. One of the examination, the Graduate Management Admission Test (GMAT), includes a critical thinking section that consists of five question types, which are "Inference," "Finding the Assumption," "Strengthening an Argument," "Weakening an Argument and Spotting the Flaws," as well as "Paradox or Discrepancy."

These question types can be further grouped as "Finding the Missing Assumption", "Strengthing an Argument / Finding a Valid Conclusion" and "Weakening an Argument / Spotting an Invalid Conclusion". Therefore, to synthesize a new dataset, we have designed ILogicEval with three similar question types. The three question types are described below, including their name and explanations.

In the question type of **3c1e**, the content part serves as a premise. It either contradicts or does not imply three of the options, while implying the remaining option. Conversely, in the question type of **3e1c**, with the content part also acting as the premise, it implies three of the options, and either contradicts or does not imply the remaining one. In the question type of **Missing Premise**, the content part is known as the premise. Subsequently, the content part is modified to combine with a valid conclusion from the entailment group. The necessary proposition in the premise, which ensures the premise implies the conclusion, is then removed. The removed proposition then becomes the correct option. The remaining three options are sampled from the non-entailment group, with additional validation carried out by an external validator. Alternatively, if there are sentences in the SNLI dataset that display a contradiction during the subsequent transformation into natural language, the three options can also be chosen from the entailment group. Sample instances of the three question types are presented in Table 1.

3.1.3 Transforming Into Natural Language

To ensure the richness of language diversity, each logical variable is replaced by simple sentences sourced from SNLI (Glockner et al., 2018) and MNLI datasets (Williams et al., 2018). Inappropriate sentences from these datasets are filtered by predefined rules to ensure language quality and details are provided in the subsequent paragraphs. For the SNLI dataset, multiple instances can have the same premise. They are grouped together, resulting in a table containing the premise, entailing hypothesis, contradicting hypothesis, and neutral hypothesis. Most of the time, only the premise and entailing hypothesis are sampled during the transformation from propositions to natural language. The contradicting hypothesis is employed when forming the three contradicting options in '3c1e' question type, the one contradicting option in '3e1c' and also the three non-ground-truth options in 'missing premise'.

To generate natural language templates for the three inference rules, ChatGPT is employed, and several sample templates are provided as input. For instance, one of the templates for expressing logical implication, "A implies B," can be used. However, when incorporating these templates into sentences that consist of multiple parts, a potential issue arises when one statement lacks a subject. In such cases, it is assumed that the subject refers to the subjects of the neighboring statements, which may introduce inconsistencies in the synthetic texts. To mitigate this issue, the Stanford POS tagger is utilized to filter out sentences beginning with the tags 'VERB' or 'AUX'. Additionally, sentences lacking a 'VERB' tag are also filtered to ensure language quality.

When handling negated variables, it is necessary to negate the corresponding sentences in natural language. The same POS tagger is employed to identify the verb and add the words "don't/doesn't/didn't" or the token "n't" in front of it. If the token "not/n't" is already present, the relevant words are reverted back to their original form.

While verbs can be identified in the remaining

instances, a notable portion of sentences in the present continuous tense lack an auxiliary verb. To address this, the nltk tagger, which offers a more detailed classification of verb tense, is utilized to reintroduce the appropriate auxiliary verb.

The finalized text is then subjected to a grammar check by passing it through ChatGPT once again. The difference in the text before and after this process is recorded. In order to measure the extent of the modifications made, we calculate the ratio of the length of the longest common substring between the original text and the modified text to the maximum length between the two texts. To ensure that ChatGPT has not made excessive changes to the content, we only keep the modifications if the computed score is greater than 0.5. Multiple trials are conducted, and the result with the smallest amount of modification is retained. In order to maintain the quality of the testing set, we manually review and approve the changes suggested by ChatGPT.

It is important to note that each instance in SNLI (after grouping the same premises) and MNLI only appears in a single instance of ILogicEval. There are no repeated uses across instances in the proposed dataset.

3.2 Dataset Overview

ILogicEval is a multiple-choice dataset, comprising four options, with one option being the correct answer. The dataset comprises a total of 12,589 instances, distributed as follows: 4196 instances corresponding to '3c1e', 4195 instances corresponding to '3e1c', and 4198 instances corresponding to 'missing premise'.

The dataset is partitioned into training, validation and testing sets. Both the validation and testing sets consist of 900 instances, with the class balance being maintained within each set. The distinct vocabulary size of ILogicEval, determined using the nltk tokenizer, is comparable to that of complex datasets such as ReClor and LogiQA. Additionally, it is significantly larger than that of first-order logic reasoning datasets. For a comprehensive overview of ILogicEval and its comparison with other datasets in the logical reasoning domain, refer to Table 5.

4 Effect in Other Logic Reasoning Task

Despite the counterintuitive nature of ILogicEval, the question types in ILogicEval are similar to those in ReClor. Previous approaches to solving the ReClor task have involved additional pretraining using extra data ((Jiao et al., 2022; Sanyal et al., 2023)). Thus, experiments were conducted to investigate whether incorporating ILogicEval as extra pretraining data for ReClor could improve task performance. The training process involved initially training LLMs using ILogicEval data, followed by ReClor data. The results obtained using the DeBERTa-v2-xlarge model as the backbone for ReClor testing are presented in Table 1. In the table, "Test" shows the overall accuracy while "Test-E" and "Test-H" are the accuracy corresponding to the EASY and HARD sets in ReClor.

	Test	Test-E	Test-H
ILogicEval	73.2	83.6	65.0
MERIt	73.1	86.2	64.4
DeBERTa-v2-xlarge	71.0	83.8	60.9

Table 1: Results with different pretraining data on ReClor with DeBERTa-v2-xlarge backbone

In addition, for the purpose of comparison, we include the results of MERIt (Yu et al., 2020), one of the state-of-the-art models that also employed an additional pretraining corpus from Wikipedia, using the DeBERTa-v2-xlarge backbone. Our results successfully outperform MERIt, demonstrating the extent to which ILogicEval can contribute to enhancing the logical reasoning capability of LLMs.

5 Performance of LLMs

When the scale of model size increases, there is an inherent capability for LLMs to handle different natural language tasks in a zero-shot setting (Wei et al., 2022; Kojima et al., 2022). In addition to studying their performance under a zero-shot setting, we further investigate their performance in a few-shot setting by providing examples for guidance. Under the three-shot setting, the models were provided with three specific examples to facilitate their learning process prior to answering each question. The inclusion of one example corresponding to each question type is ensured and sampled from the training set. The models under review include ChatGPT, GPT4 (OpenAI, 2023), Gemini (Team, 2023), Llama2 (Touvron et al., 2023) and Mixtral (Jiang et al., 2023). They are prompted with the instruction "You need to answer in the form of Answer: <A/B/C/D>".

A detailed comparison between different LLMs under the zero-shot and 3-shot settings is illustrated in Table 3. 'Accuracy' measures the accuracy of the original instances. 'Circular' is the circular evaluation proposed in a previous work (Liu et al., 2023), it involves creating four versions of a single instance by shifting the order of the options in a circular way. Specifically, if the original options are labeled A, B, C, and D, the four mutants created are (A,B,C,D), (B,C,D,A), (C,D,A,B), and (D,A,B,C), where the first mutant corresponds to the original instance. In this approach, an instance is considered correct only if all of its mutants with different options in different positions are answered correctly. 'PartialCircular' is a modified circular evaluation newly introduced in the following section, measuring both the accuracy and the level of confidence in answering.

Directly using the accuracy metric directly on the instances can lead to inaccurate and inconsistent evaluation. For instance, when comparing GPT4 and Gemini in a zero-shot setting, Gemini may outperform GPT4 in a single trial using the accuracy metric. However, when we consider the circular evaluation and the confidence level of the answers, GPT4 significantly outperforms Gemini, as illustrated in Table 2. Meanwhile, in circular evaluation, counting an instance as correct only when all four mutants are correct can lead to inconsistent evaluations. In case there is an incorrect answer in any of the mutants, the entire instance is considered incorrect. Additionally, LLMs may return answers in only limited choices but not all four options among several attempts, it demonstrates certain confidence in the correct option instead of returning answers in complete randomness. The confidence level also provides insight into the quality of their reasoning ability. Previous studies have also studied the occurrence of low confidence levels or inconsistent outputs in relation to hallucination phenomena (Fu et al., 2023; Manakul et al., 2023). Considering these, partialCircular evaluation is proposed.

5.1 PartialCircular (PC)

Through experiments, we observe the number of unique options returned by the LLMs and the frequency of returning the correct answer can be informative. Therefore, the computation of the correctness per instance is designed as,

$$\frac{c}{4} \cdot (1 + \sum_i p(i) \log(p(i)) \quad (3)$$

where c is the number of mutants being answered correctly, $p(i)$ is the probability of option i for $p(i) \neq 0$.

As there are four mutants (i.e. (A,B,C,D), (B,C,D,A), (C,D,A,B), (D,A,C,B)) with each option appearing once in each position, we calculated the correctness for each instance as the percentage of correct answers among the four mutants. Furthermore, the fewer unique options returned by the LLMs, the lower the level of randomness likely there. Considering the randomness in the answer selection of LLMs, an additional factor $(1 + \sum p(o) \log(p(o))$ based on Shannon entropy which measures the uncertainty level of a random variable is introduced.

During the circular evaluation, we computed the frequency of each option being selected by the LLMs. This is subsequently used to calculate the probability distribution among the four options for the entropy calculation. Since there are four possible outcomes, a logarithm with base four is adopted in the factor $(1 + \sum p(o) \log(p(o)))$. When each option was selected once, the computation value became zero. On the other hand, if only one option was selected among the four cases, we retained the original percentage correctness among the four mutants. Meanwhile, in addition to the four options provided in the dataset instances, an additional option o representing "none of the above" is included. This accounts for cases in which LLMs respond with "I do not know" or when all options are considered correct or incorrect by LLMs.

Under the variation of return from LLM, comparing the accuracy difference between different models is not effective. In table 2, we ran Gemini on ILogicEval five times, and we found difficulty in distinguishing Gemini's logical reasoning ability from other LLMs. With circular and partialcircular, the performance differences between different LLMs are more significant and consistent. To account for the variations caused by the circular metrics' hard cutoff, the coefficient of variance among the five runs was computed. If any of the four option-circulating versions was incorrect, the entire set of instances received a score of zero, resulting in the highest variations for this metric.

5.2 Human

Eight university graduate students are invited to complete 120 instances sampled from the testing set of ILogicEval, the average accuracy achieved is 40.0%. For a better understanding of the ability

	Gemini-1	Gemini-2	Gemini-3	Gemini-4	Gemini-5	CV
Accuracy	30.0	32.0	32.4	30.1	32.0	3.3
Circular	7.4	8.1	8.0	8.0	9.0	6.3
PC	17.6	18.8	18.1	18.5	19.1	3.1

Table 2: Coefficient of variance (CV) on five runs among difference evaluation metrics on Gemini Pro

Model	Settings	Metrics	Test	Test-3e1c	Test-3c1e	Test-missing
GPT4 (gpt-4-1106-preview)	0-shot	Accuracy	**32.2**	**37.2**	**33.0**	26.6
		Circular	**12.3**	**11.3**	13.3	12.3
		PC	**22.1**	**22.3**	21.1	23.0
	3-shot	Accuracy	27.2	29.7	27.3	24.7
		Circular	7.2	4.0	7.7	10.0
		PC	17.0	13.1	16.5	21.4
ChatGPT (gpt-3.5-turbo)	0-shot	Accuracy	29.6	29.7	28.0	30.9
		Circular	3.7	0.7	5.3	5.0
		PC	13.3	9.3	14.9	15.7
	3-shot	Accuracy	30.1	28.4	30.0	**31.9**
		Circular	4.6	1.0	7.0	5.7
		PC	16.0	12.0	17.9	18.1
Gemini (gemini-pro)	0-shot	Accuracy	30.0	27.8	31.1	31.3
		Circular	7.4	9.4	4.9	7.9
		PC	17.6	19.1	15.7	17.9
	3-shot	Accuracy	29.6	28.6	29.2	30.9
		Circular	8.6	8.9	9.0	8.0
		PC	17.9	17.8	18.3	17.6
Llama2 (llama-2-70b-chat-hf)	0-shot	Accuracy	26.2	27.7	26.7	24.3
		Circular	2.8	0.0	5.3	3.0
		PC	12.9	7.8	17.3	13.7
	3-shot	Accuracy	28.6	27.4	28.7	29.6
		Circular	6.0	2.0	8.3	7.7
		PC	15.9	10.7	18.1	18.8
Mixtral (mixtral-8x7B-instruct-v0.1)	0-shot	Accuracy	29.9	28.7	31.0	29.9
		Circular	8.2	5.3	11.7	7.7
		PC	16.9	11.9	20.4	18.3
	3-shot	Accuracy	30.9	34.1	30.7	**28.0**
		Circular	12.2	7.7	**14.0**	**15.0**
		PC	20.9	16.2	**22.2**	**24.3**
Human	0-shot	Accuracy	40.0	46.2	32.6	42.1

Table 3: Performance with respect to the three question types in ILogicEval under different settings.

of the interviewee to solve logic puzzles, we also invited them to finish another 120 symbolic form instances sampled from the training set of ILogicEval, they achieved an average accuracy of 46.7%. Furthermore, it was observed that the readability of the questions had a negative impact on the interviewee's motivation to complete the task during the post-event interview.

5.3 Language Model Evaluation

Evaluation is performed across five LLMs, including GPT4, ChatGPT, Gemini, Llama2 and Mixtral. The corresponding model versions used for evaluation are specified below the LLM model name in Table 3.

GPT4 exhibits superior performance compared to other LLMs in general. However, even when considering the most lenient measure, its accuracy of 32.2% indicates only some level of understanding beyond random guessing, leaving significant room for improvement. Its performance in the "3e1c" question type achieves the best performance. Mixtral under the few shot settings achieves comparable performance with GPT4 and achieves the best performance in another two question types "3c1e" and "missing premise". Gemini achieves the average performance among the five LLMs. ChatGPT is worse than the Gemini model and exhibits the poorest performance in the "3c1e" question type. Llama2 performs the worst overall and also in the '3e1c" question type. When it comes to "missing premise" questions, Llama2 performs the worst under the zero-shot setting while Gemini performs the worst under the few-shot setting.

Besides the ranking, different models also benefit to varying degrees when provided with few-shot samples.

5.4 Analysis

Surprisingly, GPT4, which achieves the highest performance, is the only model that does not benefit from in-context samples across all metrics. To understand the factor causing this, we also experiment with the symbolic version of ILogicEval, named as "s-ILogicEval". The result in table 4 shows GPT4 can indeed benefit from the in-context learning in the symbolic logical expression format, indicating the potentially severe negative effect posed by the unintuitive connection of sentences on GPT4, but not on other models. Notably, the performance on s-ILogicEval is significantly better than that on ILogicEval, as shown in the table.

		ILogicEval	s-ILogicEval
0-shot	Accuracy	32.2	39.1
	Circular	12.3	16.3
	PC	22.1	27.0
3-shot	Accuracy	27.2	38.1
	Circular	7.2	18.7
	PC	17.0	28.1

Table 4: Performance of GPT-4-turbo on ILogicEval in symbolic form and in natural language form

6 Conclusion

In this paper, we introduce ILogicEval, a novel dataset derived from ReClor, SNLI and MNLI, specifically designed to evaluate current LLMs in the domain of logical reasoning. The main objective of this dataset is to address various challenges associated with isolating specific reasoning abilities during evaluation, incorporating language diversity, and preventing data contamination. Through empirical experimentation conducted on the ReClor dataset, our result demonstrates the efficacy of ILogicEval in enhancing the model's logical reasoning capabilities. Furthermore, the experiments conducted on current LLMs reveal their limitations in effectively solving complex logical reasoning tasks, thereby highlighting the need for further improvements in this area.

To investigate the logical reasoning abilities of different popular LLMs, a comparative analysis is performed. The results indicate that GPT4 exhibits the highest performance, but struggling to learn from in-context examples. To eliminate the inaccuracy in evaluation caused by the bias and randomness of LLMs, this paper proposes a new evaluation metric based on entropy for better assessing their reasoning ability.

References

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, and et al Nicholas Joseph. 2021. Evaluating large language models trained on code.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dia-

logue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning - The Good, the Bad, and the Ugly*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

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

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

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

Yongsheng Liu, Yanxing Qi, Jiangwei Zhang, Connie Kou, and Qiaolin Chen. 2023. Mmbench: The match making benchmark. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1128–1131, New York, NY, USA. Association for Computing Machinery.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Qiang Fu, Yan Gao, Jian-Guang Lou, and Weizhu Chen. 2022. Reasoning like program executors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 761–779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Soumya Sanyal, Zeyi Liao, and Xiang Ren. 2022. RobustLR: A diagnostic benchmark for evaluating logical robustness of deductive reasoners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9614–9631, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Soumya Sanyal, Yichong Xu, Shuohang Wang, Ziyi Yang, Reid Pryzant, Wenhao Yu, Chenguang Zhu, and Xiang Ren. 2023. APOLLO: A simple approach for adaptive pretraining of language models for logical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6308–6321, Toronto, Canada. Association for Computational Linguistics.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Haoye Tian, Weiqi Lu, Tsz On Li, Xunzhu Tang, Shing-Chi Cheung, Jacques Klein, and Tegawendé F. Bissyandé. 2023. Is chatgpt the ultimate programming assistant – how far is it?

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and et al Shruti Bhosale. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

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

Xi Zhao, Tingrui Zhang, Yuxiao Lu, and Guiquan Liu. 2022. Locsgn: Logic-contrast semantic graph network for machine reading comprehension. In *Natural Language Processing and Chinese Computing*, pages 405–417, Cham. Springer International Publishing.

A Overview of ILogicEval and other logical reasoning datasets

Datasets overview is shown in Table 5.

B Algorithm of constructing symbolic logical propositions

Algorithm 1 Pseudo code of formatting the content of MCQA with symbolic logical propositions

Input: A candidate list x of 8 variables; a candidates picking counter o initialized as all 0; A predefined value n decides the maximum number of propositions being created for one instance.

Randomly pick a value l between 2 and n+1.
for $i = 0$ **do**
  if $i = l$ **then**
    Break.
  end if
  if $o_i = n - 1$ **then**
    Set $p(o_i) = 0.1$
  else if $o_i \geq n$ **then**
    Set $p(o_i) = 0$
  else
    Calculate $p(o_i)$ with Eq.(2)
  end if
  $i = i + 1$
  Sample 3 variables from x according to o.
  Sample a rule from Eq.(1c) and fit the 3 variables inside.
  Add 1 to the o_j if the variable j is fit into the rule.
end for

Output: A set of propositions q

C Examples of different question type of ILogicEval

Examples corresponding to each question type are illustrated in Table 1.

	ILogicEval	ReClor	LogiQA	RuleTaker	LogicNLI	FOLIO	RobustLR
Source	synthetic, human	exam	exam	synthetic	synthetic	synthetic, human	synthetic
# of options	4	4	4	2	4	3	3
Size	12589	6138	8678	500k	20k	1435	360k
Training set size	10789	4638	6942	350k	16k	1004	200k
Validation set size	900	500	868	50k	2000	204	40k
Testing set size	900	1000	868	100k	2000	227	120k
Vocabulary size	27466	26576	56407	101	1077	4351	46

Table 5: Overview of ILogicEval and other logical reasoning datasets

3e1c
Content:
A woman outside covers her face with fabric. The woman is protecting her face, thus a woman in a striped hoodie holds a camera on a beach. Once a cheerleader in a blue shirt performing, female cheerleader is doing a handstand on a court. A woman outside covers her face with fabric as long as a young person moves around or a boy is wearing a green power ranger costume is on a ride at an arcade. Someone is not performing. $H, (H \rightarrow A), (C \rightarrow B), ((C \vee F) \rightarrow H), \neg B$
Question: Which conclusion does not follow from the provided information?
Options:
A. In the presence that a cheerleader in a blue shirt don't performing, a boy wearing a green power ranger costume is not on a ride at an arcade. $(\neg C \rightarrow \neg F)$
B. The woman is protecting her face, once someone is performing. $(B \rightarrow H)$
C. A cheerleader in a blue shirt don't performing.$\neg C$
D. A young person moves around, hence a woman in a striped hoodie does not hold a camera on a beach. $(C \rightarrow \neg A)$
Answer: A

3c1e
Content:
As long as two young men are in a boat heading away from a larger boat with a single man on it, two little boys are running from a lake with ducks. A team of dogs pulls a sled through the snow. There are men in the water on boats. In the event that not both a team of fierce canines vigorously haul a sled through snowy conditions and a team of dogs pulls a sled through the snow, humans are running. $(D \rightarrow B), F, D, (\neg(C \wedge F) \rightarrow B)$
Question: Based on the information given, which is the most inaccurate conclusion?
Options:
A. A team of fierce canines vigorously doesn't haul a sled through snowy conditions when humans are running. $(B \rightarrow \neg C)$
B. Only if a team of dogs doesn't pull a sled through the snow, humans are running. $(B \rightarrow \neg F)$
C. Only if there are men in the water on boats, a team of dogs doesn't drag a sled through the snow. $(\neg C \rightarrow D)$
D. Some playful dogs don't chase each other in the snow, only if there are men in the water on boats. $(contradict)$
Answer: C

Missing Premise
Content:
A man in a green shirt is hailing a cab. If a man in a green shirt doesn't hail a cab or kids don't play soccer outside, a group of guys is playing soccer in a park with onlookers in pavilions behind them. If it is not the case that both a dog catches a disk in the air and a group of guys are playing soccer in a park with onlookers in pavilions behind them, kids play soccer outside. Therefore, man is getting a cab, hence an animal is jumping. $E, (\neg(E \wedge F) \rightarrow C), (\neg(D \wedge C) \rightarrow F)| = (E \rightarrow D)$
Question: What is the absent assumption that links the premises to the conclusion?
Options:
A. Only if a group of guys don't play soccer in a park with onlookers in pavilions behind them, children are engaging in outdoor sports. $(F \rightarrow \neg C)$
B. A dog catches a disk in the air. D
C. Children are engaging in outdoor sports, if men are playing soccer in a park. $(C \rightarrow F)$
D. Children are engaging in outdoor sports. F
Answer: B

Figure 1: Illustration with respect to the three question types of ILogicEval