Multi-LogiEval: Towards Evaluating Multi-Step Logical Reasoning Ability of Large Language Models

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) continue to exhibit remarkable performance in natural language understanding tasks, there is a crucial need to measure their ability for human-like multi-step logical reasoning. Existing logical reasoning evaluation benchmarks often focus primarily on simplistic single-step or multi-step reasoning with a limited set of inference rules. Furthermore, the lack of benchmarks for evaluating non-monotonic reasoning represents a crucial gap since it aligns more closely with human-like reasoning. To address these limitations, we propose *Multi-LogiEval*, a comprehensive evaluation benchmark encompassing multi-step logical reasoning with various inference rules and depths. *Multi-LogiEval* covers three logic types—proportional, first-order, and non-monotonic—consisting of more than 15 inference rules and more than 50 of their combinations. Leveraging this benchmark, we conduct evaluations on a range of LLMs such as GPT-4, ChatGPT, GPT-3, LLaMa-2, and FLAN-T5, employing a zero-shot chain-of-thought. Experimental results show that there is a significant drop in the performance of LLMs as the reasoning steps/depth increases (average accuracy of $\sim$ 43% at depth-1 to $\sim$ 22% at depth-5). We further conduct a thorough investigation of reasoning chains generated by LLMs which reveals several important findings. We believe that *Multi-LogiEval* facilitates future research for evaluating and enhancing the logical reasoning ability of LLMs[1].

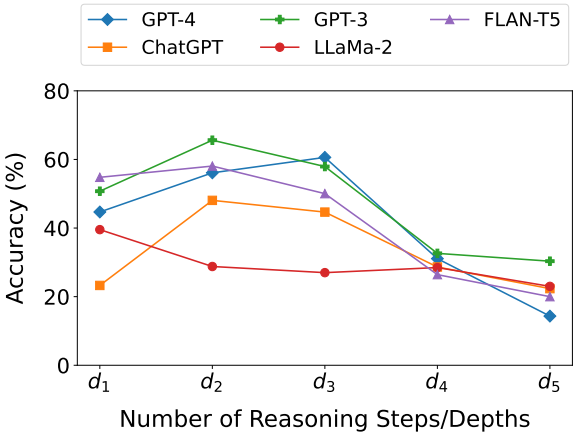

Figure 1: Performance (average accuracy across each depth) of various LLMs on *Multi-LogiEval*.

1 Introduction

The ability to perform multi-step reasoning–drawing conclusions from provided multiple premises–is a hallmark of human intelligence. Recently, Large Language Models (LLMs) such as GPT-4, GPT-3 (Brown et al., 2020b), ChatGPT, and Llama-2 (Touvron et al., 2023) have achieved impressive performance on a variety of language tasks that were previously thought to be exclusive to humans (OpenAI, 2023; Brown et al., 2020a; Zhao et al., 2023). However, the ability of these LLMs to perform multi-step logical reasoning over natural language remains under-explored, despite its various real-world applications (Khashabi, 2019; Beygi et al., 2022). Although several datasets have been proposed (Luo et al., 2023) to evaluate the logical reasoning capabilities of LLMs, these datasets are limited in their scope by (1) evaluating simplistic single-step logical reasoning such as ProntoQA (Saparov and He, 2023) and (2) evaluating multi-step logical reasoning, but only on a single type of logic and covering only a few logical inference rules as done in FOLIO (Han et al., 2022) and ProofWriter (Tafjord et al., 2021). Furthermore, there are only a few benchmarks, such as LogicBench (Paper-Under-Review, 2023) and BoardgameQA (Kazemi et al., 2023), that cover reasoning such as non-monotonic which is closer to human-like reasoning. Motivated by this, our work aims to bridge these gaps by creating a more comprehensive and logically complex evaluation

[1]Data is available at https://anonymous.4open.science/r/Multi_LogicEval-0545

benchmark where we achieve logical complexity by incorporating varying numbers of reasoning depths (i.e., multi-steps) to reach conclusions, mirroring real-world scenarios more accurately. In addition, there have been attempts made to evaluate the multi-hop reasoning of language models (Mavi et al., 2022). In contrast, our work systematically evaluates multi-hop logical reasoning over various logical inference rules and their combinations.

To this end, we propose *Multi-LogiEval*, a systematically created Question-Answering (QA) benchmark covering multi-step logical reasoning across three different logic types: Propositional Logic (PL), First-Order Logic (FOL), and Non-Monotonic (NM) reasoning and various inference rules. In particular, our proposed benchmark provides $\sim 3.5k$ instances that cover over 15 inference rules and reasoning patterns and more than 50 complex combinations of these inference rules with a different number of reasoning steps ($1 \sim 5$). To evaluate LLMs on our benchmark, we formulate a binary classification task in *Multi-LogiEval* where the context represents a natural language story consisting of logical statements, and the models have to determine whether the story logically entails a conclusion given in the question. Examples of instances are presented in Table 4. To develop *Multi-LogiEval*, we propose a two-stage procedure: (i) creating meaningful combinations of inference rules to generate data instances with different reasoning depths, and (ii) prompt LLMs to generate <*context, question, answer*> triplets consisting of different 'ontologies' (i.e., a collection of concepts such as car, person, and animals).

We evaluate a range of LLMs, including GPT-4, ChatGPT, GPT-3, Llama-2, and FLAN-T5 (Wei et al., 2021) on *Multi-LogiEval* using Zero-shot Chain-of-Thought (Zero-shot-CoT) prompting (Wei et al., 2022). The zero-shot CoT approach allows us to determine LLM's ability to do logical reasoning based on parametric knowledge (acquired during pre-training) since we can not expect in-context examples of inference rules for various reasoning depths will always be available in prompts. We measure the accuracy of LLMs' predictions on the binary classification task. As illustrated in Figure 1, our experimental results indicate that LLMs performance decreases as the depth of reasoning increases, indicating mistakes in the initial reasoning step propagate further in the reasoning chain. Furthermore, a thorough analysis of the reasoning chain generated by LLMs reveals several findings. Thus, we believe that *Multi-LogiEval* facilitates future research for evaluating the logical reasoning ability of existing and upcoming LLMs.

Dataset	Logic Covered			Multi-Step Logical Reasoning
	PL	FOL	NM	
LogicNLI	✗	✓	✗	✗
ProofWriter	✓	✓	✗	✓
FOLIO	✗	✓	✗	✓
SimpleLogic	✓	✗	✗	✓
ProntoQA	✗	✓	✗	✗
LogicBench	✓	✓	✓	✗
Multi-LogiEval	✓	✓	✓	✓

Table 1: Comparison of *Multi-LogiEval* with existing benchmarks

2 Related Work

Past attempts have been made to assess the logical reasoning ability of language models. For instance, LogiQA (Liu et al., 2021) and ReClor (Yu et al., 2020) evaluate diverse forms of logical reasoning by compiling multi-choice questions from standardized examinations, including multi-step reasoning. However, in contrast to our *Multi-LogiEval*, these datasets involve mixed forms of reasoning and do not focus on assessing logical reasoning independently. In terms of task formulation, our proposed dataset is similar to ProofWriter (Tafjord et al., 2021), FOLIO (Han et al., 2022), ProntoQA (Saparov and He, 2023), and LogicBench (Paper-Under-Review, 2023) which are QA datasets designed to evaluate logical reasoning ability independently. ProofWriter provides multi-hop proofs for each example, while FOLIO gives diverse and complex logical expressions and covers multi-step reasoning. However, it is only limited to FOL. ProntoQA (Saparov and He, 2023) provides a QA dataset with explanation and reasoning steps but is limited to single-step modus ponens in FOL. Although LogicBench (Paper-Under-Review, 2023) covers various inference rules and reasoning patterns comprehensively, it only contains single-step logical reasoning (see Table 1 for comparison). Additional datasets for evaluating multi-step logical reasoning also exist, such as SimpleLogic (Zhang et al., 2022), which only covers modus ponens inference rule, and RuleBert (Saeed et al., 2021) which covers only soft logical rules and do not evaluate logical reasoning independently. In summary, *Multi-LogiEval* evaluates logical reasoning independently and provides a multi-step logical

reasoning benchmark by creating meaningful combinations of inference rules to create data instances with different reasoning depths.

3 Multi-LogiEval

The selection of inference rules and reasoning patterns for our benchmark is motivated by LogicBench collection (Paper-Under-Review, 2023). In developing *Multi-LogiEval*, we leverage the capabilities of LLMs while employing different methods to generate data for NM compared to PL and FOL since the formulations for PL and FOL differ from NM. In particular, our data creation process consists of two stages: (i) Generation of rule combination and (ii) Generation of data instances.

Generation of rule combination We create a meaningful combination of inference rules to achieve reasoning depths and define the complex question for each combination that will require multiple reasoning steps to answer. Here, each step corresponds to one inference rule.

Generation of data instances Using the combinations of inference rules generated in the above step, we prompt the LLM to generate a more human-like story embedded with logical statements as a context and then the following complex reasoning question. In this way, we generate data in the form of <*context, question*> pairs for each combination of inference rules at each depth.

3.1 Data Generation for Monotonic Logic

Propositional Logic Propositional Logic (PL) serves as a foundational framework for reasoning about truth values of statements, represented as propositions denoted by symbols like $p, q, r, etc.$ Employing logical connectives such as '$\wedge$' (conjunction), '$\vee$' (disjunction), and '$\rightarrow$' (implication), it establishes relationships between these propositions. PL incorporates various inference rules, guiding the derivation of conclusions from given propositions. For instance, *Modus Ponens* is an example of such inference rules where if presented with the premises $((p \rightarrow q) \wedge p)$—interpreted as "if p, then q, and p is true"—we can deduce the truth of q, denoted as $((p \rightarrow q) \wedge p) \vdash q$.

First-order Logic First-order Logic (FOL) builds upon the foundations of propositional logic by introducing predicates and quantifiers. Predicates allow us to express relationships involving variables, and quantifiers such as the universal ($\forall$) and existential ($\exists$) quantifiers enable us to make statements about all or some elements in a domain. For instance, instead of stating "John is a student," we can express it in first-order logic as "There exists x such that x is John and x is a student." This logic extends the inference rules of propositional logic, such as the *Modus Ponens* rule, which lets us infer conclusions for specific instances from general premises. Here, we delve into eight distinct inference rules of PL and FOL, detailed in Table 2.

3.1.1 Generation of Rule Combination

To incorporate multi-step logical reasoning into *Multi-LogiEval*, we employ various inference rules that sequentially contribute to reaching a final conclusion as illustrated in Figure 2.

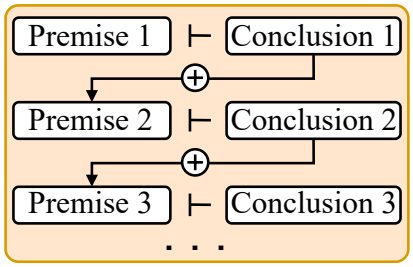

Figure 2: Process for combining multiple logical inference rules for PL and FOL. Here, *Premise 1* indicates a set of premises used for the first inference rule, and *Conclusion 1* indicates the conclusion made from these premises. *Conclusion 1* will be used along with *Premise 2* to derive *Conclusion 2*, and so on. $\vdash$: Entails.

To ensure a comprehensive approach to answering a question, we employ a method that involves leveraging both contextual information and explicit details provided in the question itself. This process requires a logical chain of reasoning, combining knowledge from the given context with the information presented in the question. Each step in this reasoning chain corresponds to a basic inference rule present in the context. We create combinations in such a way that each reasoning step corresponds to one inference rule. To generate the combinations, we start with the initial rule and assess whether the conclusion of this rule aligns with the premise of other rules. This iterative process results in multi-step combinations, with the conclusion of each step serving as a part of the premise for the subsequent rule, facilitating a layered multi-step approach to answering the question.

To explore various scenarios, we create 25 rule combinations, ranging from 2-step to 5-step reasoning chains for both PL and FOL. We use each

Rule	Propositional Logic	First-order Logic
MP	$((p \rightarrow q) \wedge p) \vdash q$	$(\forall x(p(x) \rightarrow q(x)) \wedge p(a)) \vdash q(a)$
MT	$((p \rightarrow q) \wedge \neg q) \vdash \neg p$	$(\forall x(p(x) \rightarrow q(x)) \wedge \neg q(a)) \vdash \neg p(a)$
HS	$((p \rightarrow q)) \wedge (q \rightarrow r)) \vdash (p \rightarrow r)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (q(x) \rightarrow r(x))) \vdash (p(a) \rightarrow r(a))$
DS	$((p \vee q) \wedge \neg p) \vdash q$	$(\forall x(p(x) \vee q(x)) \wedge \neg p(a)) \vdash q(a)$
CD	$((p \rightarrow q) \wedge (r \rightarrow s) \wedge (p \vee r)) \vdash (q \vee s)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (r(x) \rightarrow s(x))) \wedge (p(a) \vee r(a))) \vdash (q(a) \vee s(a))$
DD	$((p \rightarrow q) \wedge (r \rightarrow s) \wedge (\neg q \vee \neg s)) \vdash (\neg p \vee \neg r)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (r(x) \rightarrow s(x))) \wedge (\neg q(a) \vee \neg s(a))) \vdash (\neg p(a) \vee \neg r(a))$
BD	$((p \rightarrow q) \wedge (r \rightarrow s) \wedge (p \vee \neg s)) \vdash (q \vee \neg r)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (r(x) \rightarrow s(x))) \wedge (p(a) \vee \neg s(a))) \vdash (q(a) \vee \neg r(a))$
CT	$(p \vee q) \dashv\vdash (q \vee p)$	$\forall x(p(x) \vee q(x)) \dashv\vdash \forall x(q(x) \vee p(x))$

Table 2: Inference rules that establish the relationship between premises and their corresponding conclusions. MP: Modus Ponens, MT: Modus Tollens, HS: Hypothetical Syllogism, DS: Disjunctive Syllogism, CD: Constructive Dilemma, DD: Destructive Dilemma, BD: Bidirectional Dilemma, CT: Commutation.

single inference rule as depth-1. Examples of rule combinations corresponding to each depth are presented in Table 3. Let's consider a specific combination involving the *Modus Tollens* (i.e., $((p \rightarrow q) \wedge \neg q) \vdash \neg p$) and *Disjunctive Syllogism* (i.e., $((p \vee r) \wedge \neg p) \vdash r$) rules for creating combination for depth-2. Given the story in the context, including natural language statements for $(p \rightarrow q)$ and $(p \vee r)$ and information in the question as $\neg q$ in natural language, we ask about the truth value of r. Applying *Modus Tollens*, we deduce $\neg p$ from the $(p \rightarrow q)$ present in the story and $\neg q$ in the question, essentially giving the first step. Subsequently, using $\neg p$ as the premise for *Disjunctive Syllogism*, we conclude that r is indeed true based on the $(p \vee r)$ and $\neg p$, essentially giving the second step. More examples of rule combinations for each depth are given in Appendix A.

3.1.2 Generation of Data Instances

To create natural language (NL) data instances corresponding to various depths for PL and FOL, we prompt the Claude 2[2] with instructions corresponding to various rule combinations. To enhance the data generation process, we utilize a few-shot prompting. The prompt schema, as depicted in Figure 3, comprise five crucial components:

Rule Definition We manually create sets of generalized rules for various combinations, each represented by labels such as P and Q denoting propositions. For instance, consider Rule 1: "If P is true, then Q is true." Utilizing these defined rules, we construct the contextual premise by combining them. Subsequently, we formulate a question that requires a step-by-step deduction using all the established rules to derive the answer. This structured

[2] https://www.anthropic.com/index/claude-2

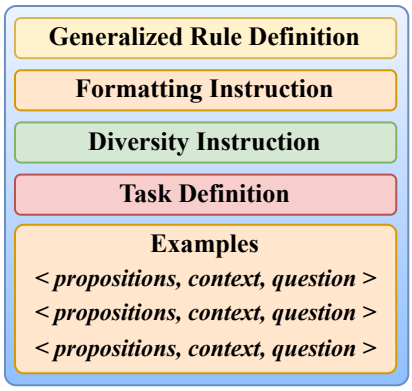

Figure 3: Schematic representation of prompt for PL. A similar structure is used for FOL.

approach allows for a comprehensive exploration of knowledge within the given context.

Format We provide the model-specific instructions for generating outputs in a designated format, simplifying the process of parsing the output on a large scale.

Introducing Diversity To enhance diversity in generated examples, we prompt the model to generate multiple instances across various domains such as education, and finance. We beforehand provide a set of diverse domains to ensure the diversity in generated instances.

Task Definitions We provide definitions to perform two tasks. First to generate the context story that serves as a human-like illustration of generalized rules. This task instructs the generation of a real-life story with sentences exemplifying the specified rules, where entity labels such as P, Q, R, S, T, and U are replaced with actual entities. To ensure clarity, entity labels are excluded from the story. Additionally, the story generation

Depth	Rule Combinations	Premises in Story	Premise in Question	Answer
1	**MT:** (P → Q) ∧ ¬Q ⊢ ¬P	(P → Q)	¬Q	**¬P:** ✓
2	**MT:** (P → Q) ∧ ¬Q ⊢ ¬P **DS:** (P ∨ R) ∧ ¬P ⊢ R	(P ∨ R), (P → Q)	¬Q	**R:** ✓
3	**HS:** (P → Q) ∧ (Q → R) ⊢ (P → R) **MP:** (P → R) ∧ P ⊢ R **MP:** (R → S) ∧ R ⊢ S	(P → Q), (Q → R), (R → S)	P	**S:** ✓
4	**CD:** (P → Q) ∧ (R → S) ∧ (P ∨ R) ⊢ (Q ∨ S) **DS:** (Q ∨ S) ∧ ¬Q ⊢ S **MP:** (S → T) ∧ S ⊢ T **MP:** (T → U) ∧ T ⊢ U	(P → Q), (R → S), (P ∨ R), (S → T), (T → U)	¬Q	**U:** ✓
5	**HS:** (P → Q) ∧ (Q → R) ⊢ (P → R) **MT:** (P → R) ∧ ¬R ⊢ ¬P **DS:** (P ∨ S) ∧ ¬P ⊢ S **MP:** (S → T) ∧ S ⊢ T **MP:** (T → U) ∧ T ⊢ U	(P → Q), (Q → R), (P ∨ S), (S → T), (T → U)	¬R	**U:** ✓

Table 3: Examples of multi-step reasoning rule combinations for PL. Similar combinations are used for FOL.

task for FOL incorporates instructions specifying the use of generalized sentences with indefinite pronouns for quantification. The second task focuses on question generation, which entails formulating questions in the format: "[If (....) is true/not true, then is (....) true?]" This dual-task approach ensures the generation of <*context, question*> pair. We provide examples of generated NL instances in Table 4 for PL and FOL.

Examples We present five varied in-context exemplars for every rule combination. Each instance comprises propositions such as P, Q, R, and more, alongside a contextual narrative and an associated question. An example prompt for depth-3 is presented in Appendix B and we follow a similar structure to create all other prompts.

3.2 Non-Monotonic Reasoning

Here, we utilize eight NM reasoning patterns defined in the Lifschitz (1989), provided in Appendix D. For NM, we only generated data for depth-1 and depth-2 of logical difficulty. We limit our data generation to depth-2 since the NM reasoning patterns presented in Lifschitz (1989) involve 4-5 assumptions for each rule, and combining two rules with classical logic results in a lengthy narrative and it becomes challenging for LLMs to generate quality instance with that long narrative. Hence, we limit NM to depth 2.

3.2.1 Generation of Rule Combination

We consider reasoning patterns corresponding to default reasoning for depth-1. We generalize the rule to generate simple sentence pairs independently before combining the template-based NM rule. After generating sentence pairs independently, we combined the sentences based on the defined rule and formulated the question-answer pair accordingly. To achieve the rules with reasoning depth-2, we combined the rules from PL and NM. We manually generate a total of 9 such rule combinations provided in Appendix D. A logical relationship between context and question is followed while formulating depth-2 rule combinations. The rule combinations include 7 rules from NM - BDR, DRI, PBD, DRO, PBD, REII, and REIII and 3 inference rules from PL - MP, MT, and DS. The overall depth-2 data is generated by establishing a logical connection between the conclusions of two NM patterns with the PL rules.

3.2.2 Generation of Data Instances

In creating prompts for data generation, we adopt a four-part structure. First, we define the task. Second, each of the three rules is explained in the prompt, representing a broad assumption and conclusion. We use X to signify the conclusion of rule 1 and Y for rule 2. Third, our method employs instruction-based generation, where we give clear instructions for creating context and questions. These instructions guide the model to adhere to specific rules, ensuring a logical connection in the process. Finally, we establish formatting guidelines for generation, ensuring a systematic model output. Appendix D shows an example of prompt.

3.3 Qualitative Analysis

We conducted a qualitative analysis before including the sample in the *Multi-LogicEval* and have high-quality 3624 samples spanning three different logic types and various reasoning depths.

Rule Combination	Context and Question
Propositional Logic **Rules:** MT, DS **Propositions:** p: It rained heavily. q: The streets are flooded. r: The drainage system is working properly.	**Context:** The weatherman had predicted heavy rainfall in the city. John knew that if it rained heavily, the streets would be flooded the next day. However, the city's drainage system may or may not be working properly to channel the water. John's wife was sure that either it rained heavily, or the drainage system was working properly, or perhaps both were true. **Question:** If the streets are not flooded, does this mean the drainage system was working properly?
First-order Logic **Rules:** BD, DS **Predicates:** p: People save money. q: Achieving financial stability. r: People invest wisely. s: People avoid risky ventures.	**Context:** People believe that if they saved money diligently, they would achieve financial stability, and if they invested wisely, they would avoid risky ventures. However, circumstances can take a turn for the worse. They either save money or do not avoid risky financial ventures. **Question:** Mark didn't achieve financial stability. Can we conclude that Mark was investing wisely?
Non-monotonic Logic **NM rule:** BDR, DRI **PL rule:** MT **Logic:** Conclusion of BDR: *X*, Conclusion of DRI: *Y* MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	**Context:** John and Kate were both painters living in Florence during the Renaissance. As artists of that time, they were expected to accurately portray perspectives and proportions in their paintings. However, John's paintings lacked proper perspective. On the other hand, Kate had mastered the techniques of linear perspective, and her paintings displayed accurate depth and dimension. **Question:** Since Kate could paint using linear perspective but John could not, would it be correct to say that linear perspective was commonly mastered by Renaissance painters?

Table 4: Natural language examples of different rule combinations for all three logic types. More examples are presented in Appendix C.

Validation of Generated Data Instances To assess the quality of the generated samples, we leverage the GPT-3 (davinci-003) model to check the quality of generated samples. To utilize GPT-3, we formulate a set of binary questions related to the quality of data and prompt the model to answer. Based on all answers, we assess the quality of the sample. We have nine different sets of questions to validate the sample quality for PL, FOL, and NM. To make sure GPT-3 is evaluating correctly, we first randomly sample 250 data instances for various depths across three logic types. We prompt GPT-3 to assess the quality of these samples, then we manually evaluate the accuracy of the model. We find that $\sim 95\%$ of the data instances validated by GPT-3 is correct. More details related to validation are presented in Appendix E.

Statistics *Multi-LogicEval* has a total of 5 different logical reasoning depths. For PL and FOL, we have data with all 5 reasoning depths, while NM only has depth-1 and depth-2 logical reasoning. Table 5 shows the depth-wise statistics of samples present for each logic type after validation. Initially, we generate 50 samples corresponding to each rule combination. We only select data instances validated by GPT-3 which gives us 500, 1293, 849, 682, and 300 samples for depth-1, depth-2, depth-3, depth-4, and depth-5, respectively.

Logic	**Reasoning Depth**					**Total**
	1	2	3	4	5	
PL	160	549	449	347	150	1655
FOL	180	295	400	335	150	1360
NM	160	449	-	-	-	609
Total	500	1293	849	682	300	**3624**

Table 5: Statistics of *Multi-LogicEval*: Number of samples for each depth.

4 Results and Analysis

In this section, we present detailed information related to the experimental setup, our primary results, and a detailed analysis of the results.

4.1 Experimental Setup

Task Formulation We formulate a binary classification task using *Multi-LogiEval* to evaluate the multi-step logical reasoning ability of LLMs. *Multi-LogiEval* consists of data instances with different reasoning steps/depths, it is important to analyze the performance at each depth level. Let us consider a set of data instances $\mathcal{I}_{D,L}$ corresponding to depth D and logic type L. In this set, i^{th} instance is represented as $\mathcal{I}^i_{D,L} = \{(c_i, q_i)\}$ where

Models	Propositional					First-Order					Non-Monotonic	
	d_1	d_2	d_3	d_4	d_5	d_1	d_2	d_3	d_4	d_5	d_1	d_2
GPT-4	37.50%	58.18%	**58.20%**	31.42%	17.33%	47.78%	68.33%	**63.00%**	**30.74%**	11.33%	48.75%	41.78%
ChatGPT	22.50%	44.00%	41.80%	27.14%	28.00%	21.67%	72.67%	47.50%	30.15%	16.67%	25.63%	27.55%
GPT-3	37.50%	**63.27%**	57.60%	**36.57%**	**40.67%**	**57.22%**	**88.00%**	58.25%	28.66%	20.00%	57.50%	**45.56%**
LLaMa-2	30.00%	35.82%	29.00%	28.28%	19.33%	46.11%	41.67%	25.00%	28.67%	**26.67%**	42.50%	8.88%
FLAN-T5	**48.75%**	57.64%	47.80%	26.86%	26.00%	55.00%	79.00%	52.25%	25.97%	14.00%	**60.63%**	37.56%
Avg	35.25%	51.78%	46.88%	30.05%	26.27%	45.56%	69.93%	49.20%	28.84%	17.73%	47.00%	32.27%

Table 6: Evaluation of LLMs in terms of accuracy on *Multi-LogiEval*.

c_i represents context and q_i represents question corresponding to i^{th} instance. Each context (c) represents a story embedded with natural language logical statements, and question (q) represents the conclusion (see Table 3 for example). Here, each context and question pair is created in such a way that the conclusion provided in the question always entails context. However, you require different reasoning steps to reach to conclusion. We prompt the model to assign a label Yes if the conclusion logically entails the context; otherwise, assign a label No. To evaluate any LLMs on this setup, we provide $< p, c, q >$ as input to predict a label Yes or No where p is a natural language prompt.

> Given the context that contains rules of logical reasoning in natural language and question, perform step-by-step reasoning to answer the question. Based on context and reasoning steps answer the question ONLY in 'yes' or 'no'. Please use the below format:
> **Context:** [text with logical rules]
> **Question:** [question that is based on context]
> **Reasoning steps:** [generate step-by-step reasoning]
> **Answer:** Yes/No

Experiments We evaluate a range of prompting models (i.e., GPT-4, GPT-3 (davinci-003), ChatGPT, and LLaMa-2 (7B)), and instruction-tuned model (FLAN-T5 (3B)) on *Multi-LogiEval*. The evaluation is conducted on the versions of GPT-4, GPT-3, and ChatGPT released in November 2023. Each model is evaluated in a zero-shot setting where the chain-of-thought prompt is provided to the model without any in-context examples. The prompt used for experiments is provided above. We evaluate LLMs in a zero-shot setting to show the logical reasoning ability of the model based on parametric knowledge (knowledge acquired during pre-training) since we can not expect in-context examples corresponding to different reasoning patterns and depths during inference.

Metrics Here, we evaluate performance in terms of accuracy. Since the objective is to assess the model's ability to arrive at the correct conclusion, we measure the accuracy associated with the model's generation of a Yes and No label based on answer of given question.

4.2 Main Results

Table 6 illustrates the accuracy of reasoning at different depths for various Logical Language Models (LLMs), offering significant insights into their performance across distinct logic types and depths. From Table 6, experimental results reveal a consistent trend across PL and FOL, i.e., as the reasoning depth increases from 1 to 5, the models' performance drops. In particular, at depths 4 and 5, accuracy drops to $\sim 25\%$ for the majority of LLMs we evaluated. For instance, the accuracy of GPT-4 demonstrates a substantial drop from 37.50% at depth d_1 to 17.33% at depth d_5 for PL, indicating the challenge encountered even by larger-scale LLMs like GPT-4 when handling longer chains of logical reasoning. Moving on to NM, going from d_1 to d_2, there is a decrease in the performance of LLMs from an average of 47.00% to 32.27%. This suggests that combining even two non-monotonic reasoning patterns increases the difficulty for models in carrying out logical reasoning. While these models display competitive performance for d_1, d_2, and d_3, there is a significant drop in the performance of LLMs for d_4 and d_5 in the majority of cases.

From Table 6, we can observe that GPT-3 shows an average superior performance across the table compared to other LLMs. Notably, ChatGPT exhibits a comparatively lower accuracy of 22.50% even at reasoning depth d_1. Whereas FLAN-T5 achieves an accuracy of 48.75% at reasoning depth d_1, surpassing larger models such as GPT-4 and GPT-3 in this specific scenario. Furthermore, the performance of LLaMa-2 decreases from 42.50%

to 8.88% for NM. In addition, models struggle more with inference rules of PL and FOL than NM reasoning. In the below section, we discuss these findings in detail, as it is crucial to understand limitations of LLMs in carrying out logical reasoning.

4.3 Analysis and Discussion

In this section, we manually analyze the generated reasoning chains [3] by different LLMs and investigate the above-mentioned findings in detail.

Lower performance of GPT-4 and ChatGPT *vs.* GPT-3 The study compared the performance of GPT-3, GPT-4, and ChatGPT in logical reasoning tasks, focusing on various depths of reasoning. We randomly selected five samples each for PL, FOL, and NM and for every depth to assess the models' abilities. The analysis revealed a consistent trend: GPT-3 demonstrated effective capture of logical rules in the initial reasoning step, enhancing its predictive accuracy. Specifically, GPT-3 excelled in reaching conclusions for reasoning depths 4 and 5, where longer chains of accurate reasoning were required. In contrast, larger models like ChatGPT and GPT-4 exhibited a more generalized initial reasoning chain, posing challenges for accurate reasoning within a given context. The study suggests that GPT-4 and ChatGPT showed lower performance, indicating potential issues with overfitting that impacted their reasoning capabilities. This underscores the importance of refining GPT-4 and ChatGPT to improve their logical reasoning capacities, addressing the observed performance gap compared to GPT-3. Further enhancements in these models are crucial for advancing their effectiveness in logical reasoning tasks.

Why ChatGPT performance is lower on d_1? In this study, we randomly selected five samples each for depth-1 of PL, FOL, and NM logic to evaluate ChatGPT's performance. Upon analyzing the reasoning chains of these samples, we observed that the initial reasoning steps were quite generic, allowing for numerous possibilities throughout the entire reasoning chain. The limited context size played a role in ChatGPT's behavior at depth-1, resulting in inaccurate predictions. However, as the depth increased, enabling a longer context, the accuracy of predictions improved. Nevertheless, at depths 4 and 5, the precise logical reasoning posed challenges, impacting performance compared to depths 1, 2, and 3. The heightened depth introduced difficulties in maintaining precise logical connections. Conversely, at depths 4 and 5, the excess information to process hindered ChatGPT's logical reasoning capabilities, leading to an overall decrease in performance. These findings underscore the importance of balancing context length and depth for optimal performance in natural language processing tasks.

Performance of FLAN-T5 *vs.* GPT-family Models When comparing FLAN-T5 and GPT-family models, FLAN-T5 performs well in simpler reasoning tasks at depths 1, 2, and 3. In comparison to GPT family models, FLAN-T5 outperforms ChatGPT in PL at depths 1, 2, and 3, and it also surpasses GPT-4 in FOL for depths 1 and 2. However, we observed that FLAN-T5 encounters difficulties in maintaining correct logical connections as the task complexity increases. This limitation explains why FLAN-T5 excels in depth-1 tasks involving simple reasoning. In contrast, GPT-family models excel in preserving accurate logical connections, resulting in better performance at higher depths. Notably, in NM and PL at depth 1, FLAN-T5 outperforms all GPT-family models. Nevertheless, its performance diminishes when handling the combination of PL and NM logic. This analysis highlights that smaller models like FLAN can excel in reasoning for lower depths, but their understanding of higher-depth reasoning is limited.

5 Conclusions

In this work, we introduced *Multi-LogiEval*, a comprehensive multi-step logical reasoning benchmark consisting of three types of logic and over 50 combinations of inference rules. Our approach utilized two stage methodology to construct data instances for our benchmark consists of $\sim$ $3.5k$ data instances with $1 \sim 5$ reasoning depth. We evaluated a range of LLMs including GPT-4, ChatGPT, GPT-3, LLaMa-2, and FLAN-T5 on *Multi-LogiEval*. Experimental results revealed that these models struggle on performing logical reasoning, and their performance drops as the depth of logical reasoning increases (average accuracy of $\sim 43\%$ at depth-1 to $\sim 22\%$ at depth-5). Furthermore, we analyzed the reasoning chain generated by LLMs at various depth and presented interesting findings. We hope that *Multi-LogiEval* will facilitate the future research in evaluating and enhancing ability of existing and upcoming LLMs for logical reasoning.

[3] https://anonymous.4open.science/r/Multi_LogicEval-0545/reasoning_results/

Limitations

Though *Multi-LogiEval* facilitates the evaluation of the multi-step logical reasoning ability of LLMs, the complexity of reasoning depth presented in *Multi-LogiEval* can be improved by adding reasoning depth beyond 5 steps. *Multi-LogiEval* can be further extended by incorporating other inference rules and logic types. We also note that this research is limited to the English language and can be extended to multilingual scenarios for evaluating logical reasoning ability of LLMs.

Ethics Statement

We have used AI assistants (Grammarly and ChatGPT) to address the grammatical errors and rephrase the sentences.

References

Sajjad Beygi, Maryam Fazel-Zarandi, Alessandra Cervone, Prakash Krishnan, and Siddhartha Jonnalagadda. 2022. Logical reasoning for task oriented dialogue systems. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 68–79, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information. *arXiv preprint arXiv:2306.07934*.

Daniel Khashabi. 2019. *Reasoning-Driven Question-Answering for Natural Language Understanding*. University of Pennsylvania.

Vladimir Lifschitz. 1989. Benchmark problems for formal nonmonotonic reasoning: Version 2.00. In *Non-Monotonic Reasoning: 2nd International Workshop Grassau, FRG, June 13–15, 1988 Proceedings 2*, pages 202–219. Springer.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

Man Luo, Shrinidhi Kumbhar, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, Chitta Baral, et al. 2023. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. *arXiv preprint arXiv:2310.00836*.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation. *arXiv preprint arXiv:2204.09140*.

OpenAI. 2023. Gpt-4 technical report.

Paper-Under-Review. 2023. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Submitted to The Twelfth International Conference on Learning Representations*. Under review.

Mohammed Saeed, Naser Ahmadi, Preslav Nakov, and Paolo Papotti. 2021. RuleBERT: Teaching soft rules to pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1460–1476, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Combinations of rules

We created 25 multi-step reasoning inference rule combinations for Propositional Logic (PL), with depths ranging from 2 to 5. We use the same rule combinations for First Order Logic (FOL) for each depth. All rule combinations for 2-step, 3-step, 4-step, and 5-step reasoning for PL and FOL are presented in Tables 7, 8, 9, and 10 respectively. For each combination, we provide the inference rules to be used for reasoning, the premises present in the context and in the question, and the complex reasoning question-answer pair.

B Example of Prompt

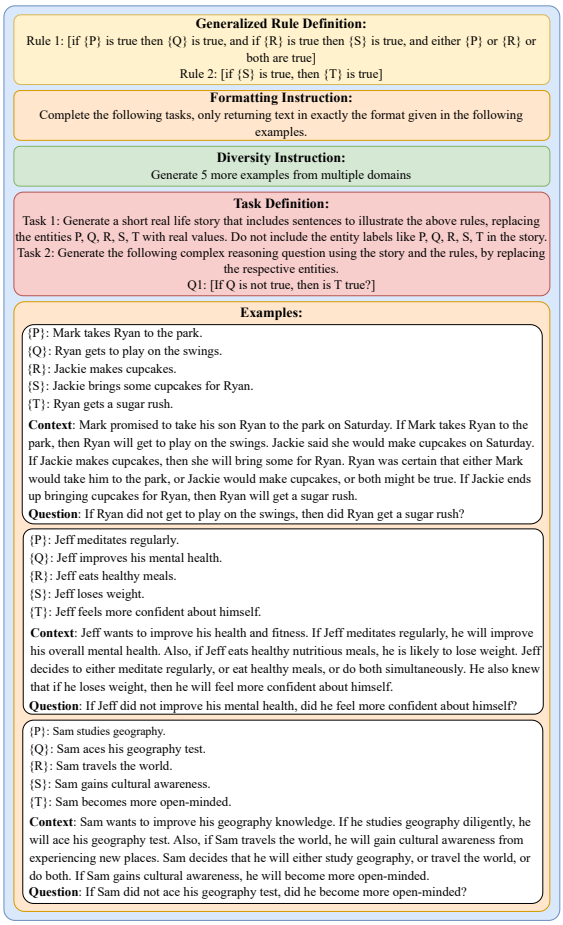

Figure 4: An example prompt for 3-step combination of inference rules CD, DS, and MP from propositional logic.

Figure 4 illustrates an example prompt for 3 depth combination of inference rules from propositional logic, namely 'constructive dilemma' (CD), 'disjunctive syllogism' (DS), and 'modus ponens' (MP). CD is formally represented as $(p \rightarrow q) \wedge (r \rightarrow s) \wedge (p \vee r)) \vdash (q \vee s)$, which can be understood in natural language as "If p implies q, and if r implies s, and either p or r or both are true, then we can conclude that either q or s or both are true." DS is formally represented as $(p \vee q) \wedge \neg p) \vdash q$, which can be understood in natural language as "If p or q are true, and we know $\neg p$, then we can conclude q." MP is formally represented as $(p \rightarrow q) \wedge p) \vdash q$, which can be understood in natural language as "If p implies q, and we know p, then we can conclude q."

In this prompt, the generalized rule definitions provide a description of the premises given in the story in natural language. The prompt includes instructions on how the generated samples should be formatted, instructions to generate samples from diverse domains, as well as detailed task definitions for generating propositions, and then using them to generate a context and question for each sample. To enhance the quality of samples in terms of relevance and coherence, the prompt includes an examples section that demonstrates these tasks. In Figure 4, we present three examples along with their respective propositions, context and question.

C NL Examples for PL and FOL

In this section, we illustrate multi-step reasoning for PL, and FOL using natural language examples for depths 2 through 5. Table 11 provides examples in natural language for PL. We provide one example of rule combinations for each depth. For each example, we provide the inference rules, and propositions, as well as the respective context and complex reasoning question. Table 12 provides examples in natural language for FOL, with one combination for each depth. Similar to PL, we provide the inference rules, predicates, and the context-question pair for each example.

D More Details on NM

Table 13 displays instances of general rules discussed in the paper by Lifschitz (Lifschitz, 1989), specifically chosen for depth-1 non-monotonic logic. Out of the 11 default non-classical reasoning rules mentioned in the paper, we opted for 8. These include Default Reasoning with Several Defaults (DRS), Default Reasoning with Irrelevant Information (DRI), Default Reasoning with a Disabled Default (DRD), Default Reasoning in an Open Domain (DRO), Reasoning about Unknown Expectations I (RE1), Reasoning about Unknown Expectations II (RE2), Reasoning about Unknown Expectations III (RE3), and Reasoning about Pri-

Rule Combinations	Premises in Story	Premise in Question	Answer
DS: $(P \lor Q) \land \neg P \vdash Q$ **MP:** $(Q \to R) \land Q \vdash R$	$(P \lor Q)$, $(Q \to R)$	$\neg P$	**R:** ✓
MT: $(P \to Q) \land \neg Q \vdash \neg P$ **DS:** $(P \lor R) \land \neg P \vdash R$	$(P \to Q)$, $(P \lor R)$	$\neg Q$	**R:** ✓
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **MP:** $(P \to R) \land P \vdash R$	$(P \to Q)$, $(Q \to R)$	P	**R:** ✓
CD: $(P \to Q) \land (R \to S) \land (P \lor R) \vdash (Q \lor S)$ **DS:** $(Q \lor S) \land \neg Q \vdash S$	$(P \to Q)$, $(R \to S)$, $(P \lor R)$	$\neg Q$	**S:** ✓
DD: $(P \to Q) \land (R \to S) \land (\neg Q \lor \neg S) \vdash (\neg P \lor \neg R)$ **DS:** $(\neg P \lor \neg R) \land P \vdash \neg R$	$(P \to Q)$, $(R \to S)$, $(\neg Q \lor \neg S)$	P	**R:** ✗
BD: $(P \to Q) \land (R \to S) \land (P \lor \neg S) \vdash (Q \lor \neg R)$ **DS:** $(Q \lor \neg R) \land \neg Q \vdash \neg R$	$(P \to Q)$, $(R \to S)$, $(P \lor \neg S)$	$\neg Q$	**R:** ✗
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **MT:** $(P \to R) \land \neg R \vdash \neg P$	$(P \to Q)$, $(Q \to R)$	$\neg R$	**P:** ✗

Table 7: 2-step reasoning rule combinations for PL and FOL.

Rule Combinations	Premises in Story	Premise in Question	Answer
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **MP:** $(P \to R) \land P \vdash R$ **MP:** $(R \to S) \land R \vdash S$	$(P \to Q)$, $(Q \to R)$, $(R \to S)$	P	**S:** ✓
CD: $(P \to Q) \land (R \to S) \land (P \lor R) \vdash (Q \lor S)$ **DS:** $(Q \lor S) \land \neg Q \vdash S$ **MP:** $(S \to T) \land S \vdash T$	$(P \to Q)$, $(R \to S)$, $(P \lor R)$, $(S \to T)$	$\neg Q$	**T:** ✓
BD: $(P \to Q) \land (R \to S) \land (P \lor \neg S) \vdash (Q \lor \neg R)$ **CT:** $(Q \lor \neg R) \dashv\vdash (\neg R \lor Q)$ **DS:** $(\neg R \lor Q) \land R \vdash Q$	$(P \to Q)$, $(R \to S)$, $(P \lor \neg S)$	R	**Q:** ✓
BD: $(P \to Q) \land (R \to S) \land (P \lor \neg S) \vdash (Q \lor \neg R)$ **DS:** $(Q \lor \neg R) \land \neg Q \vdash \neg R$ **MT:** $(T \to R) \land \neg R \vdash \neg T$	$(P \to Q)$, $(R \to S)$, $(P \lor \neg S)$, $(T \to R)$	$\neg Q$	**T:** ✗
CD: $(P \to Q) \land (R \to S) \land (P \lor R) \vdash (Q \lor S)$ **CT:** $(Q \lor S) \dashv\vdash (S \lor Q)$ **DS:** $(S \lor Q) \land \neg S \vdash Q$	$(P \to Q)$, $(R \to S)$, $(P \lor R)$	$\neg S$	**Q:** ✓
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **CD:** $(P \to R) \land (S \to T) \land (P \lor S) \vdash (R \lor T)$ **DS:** $(R \lor T) \land \neg R \vdash T$	$(P \to Q)$, $(Q \to R)$, $(S \to T)$, $(P \lor S)$	$\neg R$	**T:** ✓
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **MT:** $(P \to R) \land \neg R \vdash \neg P$ **DS:** $(P \lor S) \land \neg P \vdash S$	$(P \to Q)$, $(Q \to R)$, $(P \lor S)$	$\neg R$	**S:** ✓
DD: $(P \to Q) \land (R \to S) \land (\neg Q \lor \neg S) \vdash (\neg P \lor \neg R)$ **DS:** $(\neg P \lor \neg R) \land P \vdash \neg R$ **MT:** $(T \to R) \land \neg R \vdash \neg T$	$(P \to Q)$, $(R \to S)$, $(\neg Q \lor \neg S)$, $(T \to R)$	P	**T:** ✗

Table 8: 3-step reasoning rule combinations for PL and FOL.

Rule Combinations	Premises in Story	Premise in Question	Answer
CD: $(P \to Q) \land (R \to S) \land (P \lor R) \vdash (Q \lor S)$ **DS:** $(Q \lor S) \land \neg Q \vdash S$ **MP:** $(S \to T) \land S \vdash T$ **MP:** $(T \to U) \land T \vdash U$	$(P \to Q)$, $(R \to S)$, $(P \lor R)$, $(S \to T)$, $(T \to U)$	$\neg Q$	**U:** ✓
BD: $(P \to Q) \land (R \to S) \land (P \lor \neg S) \vdash (Q \lor \neg R)$ **CT:** $(Q \lor \neg R) \dashv\vdash (\neg R \lor Q)$ **DS:** $(\neg R \lor Q) \land R \vdash Q$ **MP:** $(Q \to T) \land Q \vdash T$	$(P \to Q)$, $(R \to S)$, $(P \lor \neg S)$, $(Q \to T)$	R	**T:** ✓
BD: $(P \to Q) \land (R \to S) \land (P \lor \neg S) \vdash (Q \lor \neg R)$ **DS:** $(Q \lor \neg R) \land \neg Q \vdash \neg R$ **MT:** $(T \to R) \land \neg R \vdash \neg T$ **DS:** $(T \lor U) \land \neg T \vdash U$	$(P \to Q)$, $(R \to S)$, $(P \lor \neg S)$, $(T \to R)$, $(T \lor U)$	$\neg Q$	**U:** ✓
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **CD:** $(P \to R) \land (S \to T) \land (P \lor S) \vdash (R \lor T)$ **DS:** $(R \lor T) \land \neg R \vdash T$ **MP:** $(T \to U) \land T \vdash U$	$(P \to Q)$, $(Q \to R)$, $(S \to T)$, $(P \lor S)$, $(T \to U)$	$\neg R$	**U:** ✓
CD: $(P \to Q) \land (R \to S) \land (P \lor R) \vdash (Q \lor S)$ **CT:** $(Q \lor S) \dashv\vdash (S \lor Q)$ **DS:** $(S \lor Q) \land \neg S \vdash Q$ **MP:** $(Q \to T) \land Q \vdash T$	$(P \to Q)$, $(R \to S)$, $(P \lor R)$, $(Q \to T)$	$\neg S$	**T:** ✓
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **MT:** $(P \to R) \land \neg R \vdash \neg P$ **DS:** $(P \lor S) \land \neg P \vdash S$ **MP:** $(S \to T) \land S \vdash T$	$(P \to Q)$, $(Q \to R)$, $(P \lor S)$, $(S \to T)$	$\neg R$	**T:** ✓
BD: $(P \to Q) \land (R \to S) \land (P \lor \neg S) \vdash (Q \lor \neg R)$ **DS:** $(Q \lor \neg R) \land \neg Q \vdash \neg R$ **MT:** $(T \to R) \land \neg R \vdash \neg T$ **MT:** $(U \to T) \land \neg T \vdash \neg U$	$(P \to Q)$, $(R \to S)$, $(P \lor \neg S)$, $(T \to R)$, $(U \to T)$	$\neg Q$	**U:** ✗

Table 9: 4-step reasoning rule combinations for PL and FOL.

Rule Combinations	Premises in Story	Premise in Question	Answer
HS: $(P \to Q) \land (Q \to R) \vdash (P \to R)$ **MT:** $(P \to R) \land \neg R \vdash \neg P$ **DS:** $(P \lor S) \land \neg P \vdash S$ **MP:** $(S \to T) \land S \vdash T$ **MP:** $(T \to U) \land T \vdash U$	$(P \to Q)$, $(Q \to R)$, $(P \lor S)$, $(S \to T)$, $(T \to U)$	$\neg R$	**U:** ✓
BD: $(P \to Q) \land (R \to S) \land (P \lor \neg S) \vdash (Q \lor \neg R)$ **CT:** $(Q \lor \neg R) \dashv\vdash (\neg R \lor Q)$ **DS:** $(\neg R \lor Q) \land R \vdash Q$ **MP:** $(Q \to T) \land Q \vdash T$ **MP:** $(T \to U) \land T \vdash U$	$(P \to Q)$, $(R \to S)$, $(P \lor \neg S)$, $(Q \to T)$, $(T \to U)$	R	**U:** ✓
CD: $(P \to Q) \land (R \to S) \land (P \lor R) \vdash (Q \lor S)$ **CT:** $(Q \lor S) \dashv\vdash (S \lor Q)$ **DS:** $(S \lor Q) \land \neg S \vdash Q$ **MP:** $(Q \to T) \land Q \vdash T$ **MP:** $(T \to U) \land T \vdash U$	$(P \to Q)$, $(R \to S)$, $(P \lor R)$, $(Q \to T)$, $(T \to U)$	$\neg S$	**U:** ✓

Table 10: 5-step reasoning rule combinations for PL and FOL.

Depth	Rules and Propositions	Context and Question
2	**Rules:** CD, DS **Propositions:** P: I practice my speech a lot. Q: I give a good presentation. R: I feel very nervous speaking in public. S: I stumble over my words during the speech.	**Context:** If I practice my speech a lot, I will give a good presentation. But if I feel very nervous speaking in public, I may stumble over my words. It seems likely either I'll practice a lot, or feel nervous, or perhaps I will practice a lot and also feel nervous before the big speech. **Question:** If I did not give a good presentation, then did I stumble over my words?
3	**Rules:** BD, DS, MT **Propositions:** P: It is raining outside. Q: The grass is wet. R: John went for a walk. S: John brought an umbrella with him. T: John has a lot of energy.	**Context:** It was a cloudy morning. Susan knew that if it is raining outside, then the grass in the yard is wet. Meanwhile, John contemplated whether to go for a walk. If John decided to go for a walk, then he brought an umbrella with him. Susan is sure that either it is raining outside, or John did not bring an umbrella with him, or it is raining outside and John did not bring his umbrella with him. She also knows that if John had a lot of energy that morning, then he went for a walk. **Question:** If the grass is not wet, then does John have a lot of energy?
4	**Rules:** HS, CD, DS, MP **Propositions:** P: It is hot outside. Q: The ice cream melts quickly. R: There are kids with sticky hands. S: The pool was crowded. T: There were long lines for the slides. U: People get frustrated and leave.	**Context:** It was the first really hot day of summer in the neighborhood. If it is hot outside, ice cream melts quickly. If ice cream melts quickly, there are kids walking around with sticky hands. The community pool also gets very crowded when it is hot. If the pool was crowded today, there were long lines for the slides. Today, either it was hot outside, or the pool was crowded, or both were true. The pool coordinator did not like this situation, because if there are long lines for the slides, then many people get frustrated and leave. **Question:** If there were no kids with sticky hands, did people get frustrated and leave?
5	**Rules:** HS, MT, DS, MP, MP **Propositions:** P: Lucy studies programming. Q: Lucy gains coding skills. R: Lucy is able to build a website. S: Lucy plays video games. T: Lucy enjoys gaming competitions. U: Lucy practices and hones her gaming skills regularly.	**Context:** Lucy wanted to start doing freelance web development work. She realized that if she studied programming, she would gain valuable coding skills. And if she has these new skills, Lucy is able to build a website on her own. As Lucy delved into her programming studies, she found herself spending hours practicing coding exercises and exploring various programming languages. In her free time, Lucy discovered a love for playing video games, finding them to be a relaxing way to unwind. It became clear that either Lucy was immersed in studying programming, or she was happily engaged in playing video games, and maybe both were true. If Lucy plays video games, it means that she enjoys gaming competitions. In her pursuit of gaming excellence, if Lucy genuinely enjoys gaming competitions, then she consistently practices and hones her gaming skills. **Question:** If Lucy is not able to build a website, then does she practice and hones her gaming skills regularly?

Table 11: Natural language examples of rule combinations of each depth for PL.

orities (RAP). These rules constitute our selection for depth-1 non-monotonic logical reasoning. Moving on to depth-2, we integrated classical and non-classical logic. Table 14 outlines the combinations of rules prepared for the depth-2 logical reasoning task. In this context, we combined BDR, DRD, DRI, PBD, DRO, REII, and REIII from non-monotonic logic with MP, MT, and DS from propositional logic to form combinations for depth-2. Table 15 shows a prompt that we have used to generate data instances for depth-2. The table shows an example of the BDR, and DRD non-monotonic logic combined with the propositional logic - DS to generate depth-2 data. The instruction-based data generation can be seen in Table 15.

E Validating Multi-LogiEval

In our research, we employed a set of validation questions, as illustrated in Table 16. A total of 9 questions were utilized to validate the generated samples. These questions were thoughtfully designed based on specific categories. To ensure comprehensive evaluation, separate sets of questions were created for Propositional logic (PL), First-Order logic (FOL), and Non-monotonic logic (NM). This approach enabled us to tailor the validation process to the unique characteristics of each rule combination within PL, FOL, and NM, thereby enhancing the accuracy and relevance of our assessment.

The PL validation question plays a crucial role in assessing the quality of a story. It ensures adherence to established rules, prevents the presence of generalized values in entities, evaluates whether the questions reflect the true logical meaning, examines the logical soundness of the story, verifies the ability to draw logical conclusions, assesses the clarity of language, and ensures logical inferences from the rules or story. In essence, it serves as a comprehensive tool to gauge and maintain the overall quality and coherence of the narrative.

Similarly, when validating First Order Logic

Depth	Rules and Predicates	Context and Question
2	**Rules:** CD, DS **Predicates:** P: Dining at an upscale restaurant Q: Enjoying a luxurious evening R: Saving money diligently S: Achieving financial stability	**Context:** One evening, someone decided to dine at an upscale restaurant. They knew that if they did, they would either enjoy a luxurious evening, if they saved money diligently, they would achieve financial stability. Someone can either dine at an upscale restaurant or save money diligently; Therefore, they will either enjoy a luxurious evening or they will achieve financial stability. **Question:** Given that Emily did not enjoy a luxurious evening. Is it true that Emily achieved financial stability?
3	**Rules:** DD, DS, MT **Predicates:** P: Eating vegetables Q: Being healthy R: Exercising regularly S: Being fit T: Being vegan	**Context:** Once upon a time, in a small town, someone decided to lead a healthy lifestyle. They knew that if they ate vegetables, they would be healthy, and if they exercised regularly, they would be fit. But not everyone in the town was healthy or fit. If they were vegan, they would eat vegetables. **Question:** We know that John exercises regularly; from the context, is it true that John is vegan?
4	**Rules:** BD, DS, MT, DS **Predicates:** P: The bus is running late. Q: Will be late for work. R: There is traffic on the road. S: Commute is longer than usual. T: It rained heavily last night. U: Will not get breakfast in the office today.	**Context:** If the bus is running late, then people will be late for work. There was also the chance of traffic on the road. If there is traffic, then people's commute takes longer than usual. Either the bus is running late, or someone's commute did not take longer than usual, or the bus is late, and someone's commute did not take extra time. If it rained heavily last night, then there is traffic on the road today. Either it rained heavily last night, or they won't get breakfast in the office this morning, or maybe both would happen. **Question:** If Mary is not late for work, then did she get breakfast in the office today?
5	**Rules:** HS, MT, DS, MP, MP **Predicates:** P: Water plants daily. Q: The plants grow bigger. R: The plants need more sunlight. S: The rainy season is near. T: Umbrellas are in high demand. U: People are preparing for inclement weather.	**Context:** Someone started growing plants as a hobby. They learned that if they water their plants daily, they will grow bigger. However, as the plants grew bigger, they needed more sunlight to thrive. But after a while, the neighbors noticed that the plants did not need more sunlight. Either the plants are watered daily, or the rainy season is near, or both. And, If the rainy season is near, then umbrellas are in high demand. It is clear that if umbrellas are in high demand, then people are surely preparing for inclement weather. **Question:** If the plants in Jim's garden do not need more sunlight, then are people preparing for inclement weather?

Table 12: Natural language examples of rule combinations of each depth for FOL.

samples, the assessment focuses on ensuring the quality of the narrative. This involves confirming the use of indefinite pronouns and appropriate pronouns for all elements outlined in the rules. Additionally, the evaluation checks for logical coherence within the story, ensuring consistency and logical connections. The assessment examines if the question is relevant to the context, clear, and connected to the story. Furthermore, the language is scrutinized for clarity, conciseness, and lack of ambiguity, and the question is assessed for its adherence to logical inferences derived from the narrative.

In order to authenticate NM samples, questions are crafted to maintain sample quality. These questions assess whether the narrative adheres to logical Rule 1 and Rule 2, avoiding generalized statements and ensuring references to objects/properties. The questions also align with Rule 3, following classical logical principles in drawing conclusions from the context. Furthermore, the posed questions are constructed based on Rule 3, ensuring that the final answers adhere to this rule. Additionally, emphasis is placed on maintaining language clarity, conciseness, and eliminating ambiguity.

Basic Default Reasoning	**Default Reasoning with Irrelevant Information**
Context: Blocks A and B are heavy. Heavy blocks are typically located on the table. A is not on the table. Conclusion: B is on the table.	Context: Blocks A and B are heavy. Heavy blocks are typically located on the table. A is not on the table. B is red. Conclusion: B is on the table.
Default Reasoning with a Disabled Default	**Default Reasoning in an Open Domain**
Context: Block A and B are heavy Heavy blocks are normally located on the table. A is possibly an exception to this rule. Conclusion: B is on the table.	Context: Block A is heavy. Heavy blocks are normally located on the table. A is not on the table. Conclusion: All heavy blocks other than A are on the table.
Reasoning about Unknown Expectations I	**Reasoning about Unknown Expectations II**
Context: Blocks A, B, and C are heavy. Heavy blocks are normally located on the table. At least one of A, B, is not on the table. Conclusion: C is on the table. Exactly one of A, B is not on the table.	Context: Heavy blocks are normally located on the table. At least one heavy block is not on the table. Conclusion: Exactly one heavy block is not on the table.
Reasoning about Unknown Expectations III	**Reasoning about Priorities**
Context: Blocks A is heavy. Heavy blocks are normally located on the table. At least one heavy block is not on the table. Conclusion: A is on the table.	Context: Jack asserts that block A is on the table. Mary asserts that block A is not on the table. When people assert something, they are normally right. Conclusion: If Mary's evidence is more reliable than Jack's. then block A is not on the table

Table 13: Illustrative examples of non-monotonic reasoning adapted from (Lifschitz, 1989).

Rule	Examples
BDR_DRD_DS Logic: Conclusion of BDR: X, Conclusion of DRD: Y DS: $(X \vee Y) \wedge \neg X \vdash Y$	Context: There were two neighboring countries, Agraria and Borduria. Agraria was known for its fertile farmlands and agriculture, while Borduria was more industrialized. Usually countries with robust agriculture also have prosperous cottage industries. However, Agraria did not have many cottage industries despite its strong agriculture. On the other hand, Borduria, with its factories and manufacturing, had many thriving cottage industries. Question: Does Borduria have prosperous cottage industries?
BDR_DRI_MP Logic: Conclusion of BDR: X, Conclusion of DRI: Y MP: $(X \rightarrow Y) \wedge X \vdash Y$	Context: Jake and Amy are authors working on their first books. Typically, first-time authors have trouble finding a publisher. Jake did not have trouble finding a publisher for his mystery novel. Charles and Gina are also first-time authors working on their books. Usually first-time authors face challenges getting their books edited properly. Gina did not have challenges getting her cookbook edited properly. Charles wrote a biography book and had helpful feedback from his friends. Question: Did Charles have trouble finding a publisher for his biography book?
BDR_DRI_MT Logic: Conclusion of BDR: X, Conclusion of DRI: Y MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	Context: John was making lasagna for dinner. He layered the noodles, sauce, cheese and other ingredients carefully. His friend Emma was also making lasagna for her family's dinner. Emma did not precook her lasagna noodles before assembling the dish. Question: Will John's lasagna noodles be cooked properly after baking?
BDR_PBD_MP Logic: Conclusion of BDR: X, Conclusion of PBD: Y MP: $(X \rightarrow Y) \wedge X \vdash Y$	Context: Rama and Lakshmana are two brothers living in Ayodhya. Generally brothers living in Ayodhya are well-versed in Sanskrit. However, Rama is not well-versed in Sanskrit. Their teacher Vashishta asserts that Lakshmana is well-versed in Sanskrit. However, their friend Bharata asserts that Lakshmana is not well-versed in Sanskrit. Normally, when Vashishta asserts something, he is right. Also, normally when Bharata asserts something, he is right too. But Bharata's evidence seems more reliable than Vashishta's. Question: Is Lakshmana well-versed in Sanskrit?
DRI_DRO_DS Logic: Conclusion of DRI: X, Conclusion of DRO: Y DS: $(X \vee Y) \wedge \neg X \vdash Y$	Context: John and Mary were students in the same math class. Normally students who studied hard for the exam passed. John did not study hard but Mary studied very diligently every day. Anna was a student in an English class. Usually students who read all the assigned books got good grades on the essays. Anna did not read all the books. Question: Did Mary pass the math exam?
DRI_PBD_MP Logic: Conclusion of DRI: X, Conclusion of PBD: Y MP: $(X \rightarrow Y) \wedge X \vdash Y$	Context: Jennifer and Megan are both pop singers who are known for their amazing dance moves during performances. Normally pop singers who are great dancers also have a big social media following. However, while Megan has over 5 million followers, Jennifer only has about 100k. On the other hand, Jennifer was invited to be a judge on a big reality dance competition show this year. Question: Does Megan have a large social media following?
DRO_PBD_DS Logic: Conclusion of DRO: X, Conclusion of PBD: Y DS: $(X \vee Y) \wedge \neg X \vdash Y$	Context: Juan has been studying French for 2 years. Normally, students who study a language for multiple years become fluent speakers. However, Juan still struggles to speak French fluently. Maria claims Juan can read French texts well because he studies hard. But Juan's teacher says his French reading comprehension is poor and he makes many mistakes. Usually Maria's assessments are accurate, and teachers' evaluations are also typically reliable. However, the teacher has more evidence from Juan's assignments and test scores to support her view. Question: Does Juan read French well?
REII_DRO_MT Logic: Conclusion of REII: X, Conclusion of DRO: Y MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	Context: John recently joined a consulting firm that helps companies formulate business strategies. The firm usually recommends companies to enter new markets only if they have strong brand presence. John was assigned to work with a clothing brand that has stores across the country but lacks brand recognition. His manager asked him to recommend strategies to improve brand presence before entering new markets. However, John felt the clothing brand should enter a few select markets first to establish itself, even without strong brand presence currently. Question: Can you determine if John's recommended strategy follows the assumptions?
REIII_DRD_MP Logic: Conclusion of REIII: X, Conclusion of DRD: Y MP: $(X \rightarrow Y) \wedge X \vdash Y$	Context: There was a country named Agravia. They had developed a new missile system called the AGM missile. This missile had advanced guidance and propulsion technology. Normally, countries that develop advanced missile systems also develop nuclear warheads to go with them. However, Agravia claimed they were only using the missiles for defense and had no plans to develop nuclear warheads. There was another country named Baronia. They had also recently developed a missile defense shield called the BMD system. Countries that develop advanced missile defense systems normally also expand their offensive missile capabilities. However, Baronia claimed they were only installing the BMD for defense and had no plans to expand their missile arsenal. Question: Does Baronia have plans to expand their missile arsenal?

Table 14: Natural language examples of rule combinations of depth-2 for NM and PL combination.

You are excellent at understanding rules and generating story around the rules indirectly.

Rule 1:
Assumptions:
1: A and B are objects of type T and have property P.
2: Normally objects of type T with property P have property Q.
3: A does not have property Q.
Conclusion:
X: B has property Q.

Rule 2:
Assumptions:
1. C and D are objects of type S and have property I.
2. Normally objects of type S with property I have property J.
3. C might not have property J even if it has property I.
Conclusion:
Y: D has property J.

Rule 3:
Assumptions:
(X or Y) is true
X is false
Conclusion:
Y is true

Instruction to generate story and question:
1. X and Y are respectively conclusion from rule 1 and rule 2, which can be derived.
2. X and Y must logically follow assumptions defined in rule 3.
3. Generated story must use all the assumptions from rule 1 and rule 2.
4. Do not refer rule, object, or property directly in the story.
5. Question from story must follow rule 3 to derive answer
6. Question should follow assumptions of rule 3.
7. Generated story should be like real stories mentioned in student's textbook.

formatting to be followed:
1. only create a story as per given instruction and question
2. Generate story with prefix: Story and question with prefix: Question
3. Generate question in new line

Table 15: An example of prompt used to generate data instance for depth-2 for non-monotonic logic - BDR, DRD and propositional logic - DS

PL	FOL	NM
1. Does the story contain sentences that follow the structure of the mentioned rules?	1. Does the story contain sentences that follow the structure of mentioned rules?	1. Does the story contain all the assumptions from Rule 1 and Rule 2 without directly referencing rules, objects, or properties?
2. Does the story use real values for the entities (P1, P2, C1, ...)?	2. Does the story use only indefinite pronouns for the elements corresponding to all rules?	2. Are the conclusions (X and Y) logically derived based on the assumptions stated in Rule 1 and Rule 2?
3. Does the story use entity labels (P1, P2, C1, ...) within the narrative?	3. Does the story use any proper noun for the elements corresponding to all rules?	3. Does the story avoid explicitly mentioning rules, objects, or properties from the logical framework?
4. Do the reasoning questions accurately reflect the logical structure of the story and rules?	4. Are the sentences in the story logically connected to establish a causal relationship as per Rules?	4. Do the assumptions in the story logically connect to form the conclusions (X and Y)?
5. Are the sentences in the story logically connected to establish a causal relationship as per the rules?	5. Does the story maintain consistency in using indefinite pronouns throughout?	5. Does the story follow the structure of Rule 3 in presenting a situation for proper conclusion?
6. Is the conclusion drawn in the story logically connected as mentioned in the given rules?	6. Is the conclusion drawn in the story logically connected as mentioned in the given rules?	6. Is the language in the story clear and concise, avoiding unnecessary details that do not contribute to the logical framework?
7. Do the reasoning questions directly relate to the content of the story and rules?	7. Is the question clear and directly related to the content of the story?	7. Does the question derived from the story follow the structure of Rule 3 and align with the assumptions made in Rule 3?
8. Is the language in the story concise and clear, avoiding ambiguity?	8. Is the language in the story concise and clear, avoiding ambiguity?	8. Is the question referring to rules, objects, or properties and not generalized?
9. Do the reasoning questions accurately follow the logical inferences derived from the story and rules?	9. Does the question accurately follows the logical inference derived from the story?	9. Does the question accurately follows the logical inference derived from the story based on the assumptions in Rule 3?

Table 16: Validation questions used to validate samples for PL, FOL and NM.