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

Nisarg Patel* **Mohith Kulkarni*** **Mihir Parmar*** **Aashna Budhiraja**
Mutsumi Nakamura **Neeraj Varshney** **Chitta Baral**

Arizona State University, USA
{nppatel7, mkulka20, mparmar3, chitta}@asu.edu

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 datasets 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 dataset encompassing multi-step logical reasoning with various inference rules and depths. *Multi-LogiEval* covers three logic types — propositional, first-order, and non-monotonic consisting of more than 30 inference rules and more than 60 of their combinations with various depths. Leveraging this dataset, we conduct evaluations on a range of LLMs including GPT-4, ChatGPT, Gemini-Pro, Yi, Orca, and Mistral, 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 $\sim 68\%$ at depth-1 to $\sim 43\%$ 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].

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, ChatGPT, Gemini, and Mistral (Jiang 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., 2020; 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 (Parmar et al., 2024) 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 eval-

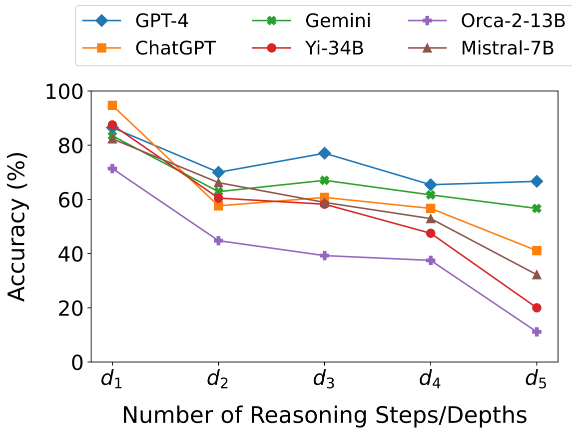

Figure 1: Performance (avg. accuracy across each depth for PL & FOL) of various LLMs on *Multi-LogiEval*.

[1]Data is available at `https://github.com/Mihir3009/Multi-LogiEval`

*Equal Contribution

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20856–20879
November 12-16, 2024 ©2024 Association for Computational Linguistics

uation dataset by incorporating varying numbers of reasoning depths (i.e., multi-steps) to reach conclusions. In addition, past attempts have been made to evaluate multi-hop reasoning of language models (Mavi et al., 2022). In contrast, our work systematically evaluates multi-hop logical reasoning over various inference rules and their combinations.

To this end, we propose *Multi-LogiEval*, a systematically created Question-Answering (QA) dataset covering multi-step logical reasoning across three different logic types: Propositional Logic (PL), First-Order Logic (FOL), and Non-Monotonic (NM) reasoning. Our objective is to present a preliminary analysis of the LLMs' ability to perform multi-step logical reasoning and demonstrate their failures even when performing simple reasoning. We believe that, regardless of whether such reasoning is available in some existing natural data (e.g., examinations), LLMs should do proper logical reasoning. Thus, we systematically compiled data using various inference rules and varying numbers of reasoning depths. In particular, our proposed dataset provides $\sim 1.6k$ high-quality instances that cover 33 inference rules and reasoning patterns and more than 60 complex combinations of these inference rules with a different number of reasoning steps ($1 \sim 5$). Our choice of inference rules is further explained in section 3.1. To evaluate LLMs, 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). In the end, we perform human validation of each generated instance to ensure the quality.

We evaluate a range of LLMs, including GPT-4, ChatGPT, Gemini-Pro, Yi-34B (Young et al., 2024), Orca-2-13B (Mitra et al., 2023), and Mistral-7B (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. The rationale behind the choice of binary classification task is that it provides systematic standard metric-based evaluation (i.e., direct comparison of LLMs' performance in terms of accuracy), which could be more challenging with open-ended question-answer formats. However, we also provide a manual and thorough analysis of the reasoning chain generated by LLMs revealing several findings such as the importance of contextual information, the lack of correlation between longer reasoning chains and better outcomes, and the lower performance of larger-scale open-source LLMs compared to smaller ones.

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

Table 1: Comparison of *Multi-LogiEval* with existing datasets and 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. Past attempts have been made to create datasets focusing on logical reasoning(Luo et al., 2023). In terms of task formulation, our proposed dataset is similar to ProofWriter (Tafjord et al., 2021), RuleTaker (Clark et al., 2021), FOLIO (Han et al., 2022), ProntoQA (Saparov and He, 2023),

Rule	Propositional Logic	First-order Logic
MP	$((p \to q) \land p) \vdash q$	$(\forall x(p(x) \to q(x)) \land p(a)) \vdash q(a)$
MT	$((p \to q) \land \neg q) \vdash \neg p$	$(\forall x(p(x) \to q(x)) \land \neg q(a)) \vdash \neg p(a)$
HS	$((p \to q)) \land (q \to r)) \vdash (p \to r)$	$(\forall x((p(x) \to q(x)) \land (q(x) \to r(x))) \vdash (p(a) \to r(a))$
DS	$((p \lor q) \land \neg p) \vdash q$	$(\forall x(p(x) \lor q(x)) \land \neg p(a)) \vdash q(a)$
CD	$((p \to q) \land (r \to s) \land (p \lor r)) \vdash (q \lor s)$	$(\forall x((p(x) \to q(x)) \land (r(x) \to s(x))) \land (p(a) \lor r(a))) \vdash (q(a) \lor s(a))$
DD	$((p \to q) \land (r \to s) \land (\neg q \lor \neg s)) \vdash (\neg p \lor \neg r)$	$(\forall x((p(x) \to q(x)) \land (r(x) \to s(x))) \land (\neg q(a) \lor \neg s(a))) \vdash (\neg p(a) \lor \neg r(a))$
BD	$((p \to q) \land (r \to s) \land (p \lor \neg s)) \vdash (q \lor \neg r)$	$(\forall x((p(x) \to q(x)) \land (r(x) \to s(x))) \land (p(a) \lor \neg s(a))) \vdash (q(a) \lor \neg r(a))$
CT	$(p \lor q) \dashv\vdash (q \lor p)$	$\forall x(p(x) \lor q(x)) \dashv\vdash \forall x(q(x) \lor p(x))$
DMT	$\neg(p \land q) \dashv\vdash \neg p \lor \neg q$	$\neg \forall x(p(x) \land q(x)) \dashv\vdash \exists x(\neg p(x) \lor \neg q(x))$
CO	$((p \to q) \land (p \to r) \vdash (p \to (q \land r)$	$\forall x((p(x) \to q(x)) \land (p(x) \to r(x))) \vdash \forall x(p(x) \to (q(x) \land r(x)))$
IM	$(p \to (q \to r)) \dashv\vdash ((p \land q) \to r)$	$\forall x(p(x) \to (q(x) \to r(x))) \dashv\vdash \forall x((p(x) \land q(x)) \to r(x))$
MI	$(p \to q) \dashv\vdash (\neg p \lor q)$	-
EG	-	$p(a) \vdash \exists x(p(x))$
UI	-	$\forall x(p(x)) \vdash p(a)$

Table 2: Inference rules that establish the relationship between premises and their corresponding conclusions. A subset of these inference rules is adapted from Parmar et al. (2024). MP: Modus Ponens, MT: Modus Tollens, HS: Hypothetical Syllogism, DS: Disjunctive Syllogism, CD: Constructive Dilemma, DD: Destructive Dilemma, BD: Bidirectional Dilemma, CT: Commutation, DMT: De Morgan's Theorem, CO: Composition, IM: Importation, MI: Material Implication, EG: Existential Generalization, UI: Universal Instantiation

and LogicBench (Parmar et al., 2024) which are QA datasets designed to evaluate logical reasoning ability independently. ProofWriter provides multi-hop proofs for each example, RuleTaker mainly covers the simple implication rules such as modus ponens, 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 (Parmar et al., 2024) 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. In contrast, *Multi-LogiEval* evaluates logical reasoning independently beyond modus ponens. In addition, FLD (Formal Logic Deduction) (Morishita et al., 2023) has formal logic theory-based inference rules, and their combinations to create multi-step reasoning, but limited to PL and FOL. However, *Multi-LogiEval* offers a broader set of inference rules for PL and FOL, along with their meaningful combinations for multi-step reasoning, in addition to NM reasoning.

3 Multi-LogiEval

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 major 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 generally 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 natural language story embedded with logical rules 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

Here, we provide details of the data generation process for PL and FOL (further details are in Appendix A). Specifically, we delve into 14 distinct inference rules of PL and FOL, detailed in Table 2.

Depth	Rule Combinations	Premises in Story	Premise in Question	Answer
1	**MT:** $(P \to Q) \land \neg Q \vdash \neg P$	$(P \to Q)$	$\neg Q$	**$\neg$P:** ✓
2	**MT:** $(P \to Q) \land \neg Q \vdash \neg P$ **DS:** $(P \lor R) \land \neg P \vdash R$	$(P \lor R)$, $(P \to Q)$	$\neg Q$	**R:** ✓
3	**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:** ✓
4	**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:** ✓
5	**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:** ✓

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

Choice of inference rules Since entailment (concluding a formula in logic from another formula in that logic) in PL is Co-NP Complete, and entailment in FOL is undecidable. Even though we are interested in multi-step reasoning, our aim is not to build a "complete" reasoning system (the system that can make all possible entailments in that logic), rather, our goal is to make LLMs be able to at least mimic some key inference rules up to a depth of five, which itself is challenging. Thus, we start with the set of 25 inference rules used in (Parmar et al., 2024) and add eight more inference rules, resulting in 33 inference rules (with zero or one variable). For a depth of five that would mean a 33^5 possible combination, which is already quite big ($>$ 39 million). In addition, we also consider seven FOL inference rules involving three variables and binary, ternary relations (Appendix H). In adding the new inference rules, our main consideration was how well they match human intuition. For example, we left out $p \land \neg p \vdash q$ as that is not very intuitive to non-logician humans. Similarly, we left out inference rules such as simplification $((p \land q) \vdash p)$, conjunction $(p, q \vdash (p \land q))$, and addition $(p \vdash (p \lor q))$, as they would lead to infinite reasoning chains and it did not make sense to add them as an additional step of reasoning to arrive at a meaningful conclusion. Conversely, we added the DMT $(\neg(p \land q) \dashv\vdash \neg p \lor \neg q)$, and show its use in multi-step, as shown in Table 9 (Appendix B).

3.1.1 Generation of Rule Combination

We apply sequential inference rules for multi-step reasoning, as illustrated in Figure 2. To ensure a comprehensive approach to answering a question,

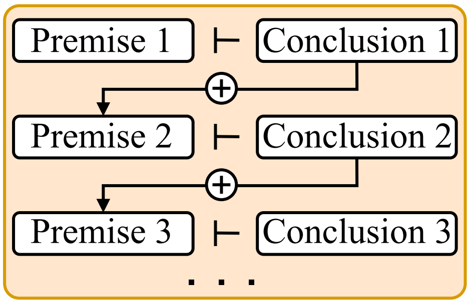

Figure 2: Process for combining multiple logical inference rules for PL and FOL: *Premise 1* is the set of premises for the first inference rule, leading to *Conclusion 1*. *Conclusion 1* and *Premise 2* derive *Conclusion 2*, and so on. $\vdash$: Entails.

we employ a method that involves leveraging contextual information and explicit details provided in the question. This process requires a logical chain of reasoning, combining knowledge from the given context with the information presented in the question. Each step in the reasoning chain corresponds to an inference rule, with combinations ensuring each step aligns with a single 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/reasoning, with the conclusion of each step serving as a part of the premise for the subsequent rule.

We create 71 rule combinations, ranging from 2-step to 5-step reasoning chains. We use each single inference rule as depth-1. Examples of rule combinations in classical logic are presented in Table 3. Let's consider a specific combination involving

Rule Combination	Context and Question
PL Rules: MT, DS **Propositions:** p: Capture shots in golden hours. q: Photo wins awards. r: Focus on rare wildlife.	**Context:** In wildlife photography, Olivia was certain that if she captured shots in the golden hours, her photos would win awards. However, opportunities varied each day. It was evident that she either captured shots during the golden hours or she focused on rare wildlife, or both. Olivia's latest photos did not win any awards. **Question:** Is it true that she focused on rare wildlife?
FOL Rules: BD, DS **Predicates:** p: Work extra hours. q: Meet project deadlines. r: Take minimal breaks. s: Increase productivity.	**Context:** In a company, employees believe that if they work extra hours, they will meet project deadlines, and if they take minimal breaks, they will increase productivity. However, they face a dilemma - they either work extra hours or do not increase productivity. **Question:** Jane didn't meet the project deadline. Is it true that Jane took minimal breaks?
NM rule: BDR **PL rule:** MP (Sentence Y) **Logic:** Conclusion of BDR: *X* MP: (X $\rightarrow$ Y) $\wedge$ X $\vdash$ Y	**Context:** Jim and Pam work at the same office. Normally, employees at that office get free lunch. Jim does not get free lunch. If Pam gets free lunch, then she gets an hour lunch break. **Question:** Can we conclude Pam gets an hour lunch break?

Table 4: NL examples of different rule combinations for all three logic types. Appendix D provides more examples.

the *Modus Tollens* ($((p \rightarrow q) \wedge \neg q) \vdash \neg p$) and *Disjunctive Syllogism* ($((p \vee r) \wedge \neg p) \vdash r$) rules for creating combination for depth-2. Given the context, including natural language statements for $(p \rightarrow q)$ and $(p \vee r)$ and information in the question as $\neg q$, we ask about the truth value of r. Applying *Modus Tollens*, we deduce $\neg p$ from the $(p \rightarrow q)$ from context and $\neg q$ in question, 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 $\neg p$, giving the second step. Creating rule combinations at higher depths, and validating the quality of generated instances is challenging, hence, we limit the number of rule combinations at d_5 for the scope of this study. More examples provided in Appendix B.

3.1.2 Generation of Data Instances

We generate natural language (NL) data at different depths by prompting Claude-2 in a few-shot setting with instructions for various rule combinations. The prompt schema, shown in Figure 3, comprise five crucial components:

Rule Definition We provide generalized rules for various combinations containing propositions represented by labels such as P and Q. For instance, 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.

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

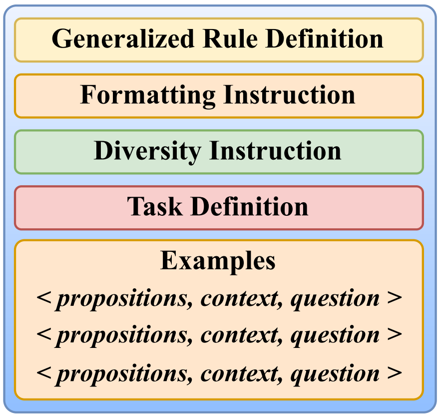

Figure 3: Data generation prompt for PL and FOL

Diversity To enhance diversity, we prompt the model to generate multiple instances across various domains, such as education and finance.

Task Definitions We provide definitions to perform two tasks. First, to generate the context 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 context. Additionally, the context generation 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: "[(....) 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.

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

3.2 Non-Monotonic Reasoning

We utilize eight NM reasoning patterns defined in Lifschitz (1989) (Appendix E), and have generated data for depths 1 to 5. To increase reasoning depth, we integrated NM with classical logic, using only one NM rule per depth due to the 4-5 assumptions each pattern involves. Thus, combining two NM patterns with classical logic creates lengthy contexts, challenging for LLMs to generate quality instances. Our rule combinations avoid overly long contexts while requiring reasoning up to depth-5.

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 them according to the template-based NM rule. After generating sentence pairs, we combined the sentences based on the defined rule and formulated the question-answer pair accordingly. We have manually generated 12, 2, 2, and 1 rule combinations for depth-2, depth-3, depth-4, and depth-5, provided in Appendix E. While formulating depth-wise rule combinations, a logical relationship between the context and question is followed. The rule combinations for all depths from 2 to 5 include 6 reasoning rules from NM—BDR, PBD, DRO, PBD, REII, and REIII—and 3 inference rules from PL—MP, MT, and DS. The data for depths 2 to 5 is generated by forming a logical connection between two NM rules' conclusions and the PL rules.

Generation of Data Instances In creating prompts for data generation, we use a four-part structure: (1) define the task, (2) explain each rule as an assumption and conclusion, (3) provide instructions for creating context and questions to ensure logical connections, and (4) establish formatting guidelines for systematic output. Appendix E shows an example of the prompt.

3.3 Qualitative Analysis

After data generation, we conducted a manual qualitative analysis, resulting in 1,552 high-quality samples for *Multi-LogicEval*.

Logic	Reasoning Depth					Total
	1	2	3	4	5	
PL	120	105	135	120	45	525
FOL	130	105	135	120	45	535
NM	160	232	40	40	20	492
Total	410	442	310	280	110	**1552**

Table 5: Statistics of *Multi-LogiEval*

Statistics *Multi-LogicEval* has 5 different logical reasoning depths. Table 5 shows the depth-wise statistics of samples present for each logic type after validation. After manual validation, from the generated data, we selected/updated high-quality 10 data instances for each inference rule in depth 1 and 15 or 20 data instances for each rule combination, which resulted in 410, 442, 310, 280, and 110 samples for depth-1, depth-2, depth-3, depth-4, and depth-5, respectively. For evaluation, of the total 1552 samples, 1126 samples have the answer *yes*, and the remaining 426 samples have the answer *no*.

Quality of Data Instances We examine each context for potential discrepancies throughout the data generation phase, ensuring they are logically correct and represent the intended logical relations. We also dedicated considerable effort to eliminating typos and validating the grammar. While validating, we encountered a few errors within the synthetically generated story-based context. We manually mitigate these errors to ensure integrity and utility (Analysis presented in Appendix F).

4 Results and Analysis

4.1 Experimental Setup

Task Formulation We formulate a binary classification task using *Multi-LogiEval*. 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 c_i represents context and q_i represents question corresponding to i^{th} instance. Here, each context and question pair is created so that the conclusion provided in the question always entails context. However, you require different reasoning steps to conclude. We prompt the model to assign a label *Yes* if the conclusion logically entails the context; otherwise, *No*. To evaluate any LLMs, we provide $< p, c, q >$ as input to predict a label *Yes* or *No* where p is a natural language prompt.

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	d_3	d_4	d_5
GPT-4	89.17	69.52	82.22	71.67	66.67	83.85	70.48	71.85	59.17	66.67	36.88	51.67	65.00	67.50	60.00
ChatGPT	91.67	56.19	63.70	62.50	44.44	97.69	59.05	57.78	50.83	37.78	33.75	41.11	50.00	62.50	60.00
Gemini	90.00	62.86	68.15	65.83	60.00	76.92	62.86	65.93	57.50	53.33	46.25	46.11	62.50	55.00	60.00
Yi-34B	85.00	65.71	58.52	46.67	26.67	90.00	55.24	57.94	48.33	13.33	37.50	41.11	55.00	62.50	65.00
Orca-13B	75.83	41.91	35.56	35.00	15.56	66.92	47.62	42.96	40.00	6.67	21.88	26.67	25.00	15.00	25.00
Mistral-7B	80.83	68.57	61.48	53.33	44.44	83.85	63.81	56.30	52.50	20.00	37.50	39.44	52.50	47.50	65.00
Avg	**85.42**	**60.79**	**61.61**	**55.83**	**42.96**	**83.21**	**59.84**	**58.79**	**51.39**	**32.96**	**35.63**	**41.02**	**51.67**	**51.67**	**55.83**

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

Experiments We evaluate a range of proprietary models (i.e., GPT-4, ChatGPT, and Gemini-Pro) and open-source models (i.e., Yi-34B-Chat, Orca-2-13B, and Mistral-7B-Instruct) on *Multi-LogiEval*. The evaluation is conducted on the versions of OpenAI and Google models released in April 2024. Each model is evaluated in a zero-shot-CoT setting. The prompt used for experiments is provided below. We evaluate LLMs in a zero-shot setting to show the logical reasoning ability of the model based on knowledge acquired during pre-training since we can not expect in-context examples corresponding to different reasoning patterns and depths during inference. However, we also evaluate LLMs in a 3-shot setting (results are in Appendix G).

Metrics Since the objective is to assess the model's ability to arrive at the correct conclusion, we measure the accuracy associated with a *Yes* and *No* label. Apart from accuracy, we provide an in-depth analysis of reasoning chains in section 4.3 to gain insights into models' performance. In addition, we would like to mention that the binary labels *Yes* and *No* indicate whether the conclusion presented in the question can be derived from the context. Hence, accuracy is an important evaluation metric, reflecting the model's reasoning ability.

> 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

4.2 Main Results

Objective Evaluation Table 6 illustrates the accuracy of reasoning at different depths for various 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' average performance drops. In particular, at depths 4 and 5, accuracy drops significantly for the majority of LLMs we evaluated. For instance, the accuracy of GPT-4, ChatGPT, and Gemini demonstrates a substantial drop from 89.17%, 91.67%, and 90% at d_1 to 66.67%, 44.44%, and 60.00% at d_5 for PL, respectively, indicating the challenge encountered even by larger-scale LLMs when handling longer chains of logical reasoning. In summary, for PL and FOL, LLMs perform well on d_1 compared to other depths. While they show competitive performance for d_2 and d_3, there is a significant drop in performance for d_4 and d_5 in most cases. In contrast, moving on to NM, going from d_1 to d_5, there is an increase in the performance of LLMs from an average of 35.63% to 55.83%.

Random Baseline We calculated a random baseline for each depth from Multi-LogiEval as below:

$$\text{Acc}_{\text{random}} = p_{\text{yes}}^2 + p_{\text{no}}^2,$$

where p_{yes} and p_{no} represent the probabilities of predicting "yes" and "no," respectively. The random baselines for depths d_1, d_2, d_3, d_4, and d_5, with corresponding $\text{Acc}_{\text{random}}$ of 86.63%, 67.35%, 53.71%, 58.33%, and 83.33%, respectively. From Table 6, we can observe that these models perform lower in terms of average accuracy compared to the random baseline.

Findings Table 6 reveal that open-source models experience a significant performance drop from d_4 to d_5. Also, there is an increasing performance trend in NM. For PL and FOL, GPT-4, ChatGPT, and Gemini show improved performance from d_2 to d_3, whereas the performance of open-source models consistently decreases. In addition, larger open-source models demonstrate decreasing performance. Furthermore, ChatGPT performs lower than GPT-4 and Gemini at d_5 in PL and FOL. Also, FOL performance is lower compared to PL at d_5.

4.3 Analysis and Discussion

In this section, we manually analyze the generated reasoning chains [2] by different LLMs and investigate the above-mentioned findings in detail.

Performance Improvement from d_2 to d_3 in PL and FOL for GPT-4, ChatGPT, and Gemini GPT-4, ChatGPT, and Gemini excel at d_3 for PL, with a performance decrease at d_4 and d_5. This trend is also observed in FOL for the same models except ChatGPT. Systematic analysis of all the reasoning chains with wrong predictions for PL and FOL shows these models reach incorrect conclusions often due to the wrong interpretation of evidence. In d_3, increasing context length improves LLMs accuracy in information mapping, thus achieving peak performance (comparison with d_2 to d_5). At d_2, around $\sim 27.4\%$ of reasoning chains with incorrect conclusions were due to the models' failure to correctly map information, either from context to conclusion or the premise from one step to the next step. This number drops to $\sim 22\%$ at d_3 and we observed that a larger context length at d_3 helps in reducing this problem. However, at d_4 and d_5, the length of the reasoning chain increases further. Since longer reasoning steps are more prone to error propagation at later stages, causing the models to deviate further from the true conclusion, hence, lower performance at d_4 and d_5.

Lower Performance of ChatGPT compared to GPT-4 and Gemini at Higher Depths This pattern is particularly evident in FOL and PL at d_5 for ChatGPT compared to Gemini, and GPT-4. At d5, manual analysis shows that ChatGPT tends to generate longer reasoning chains compared to Gemini, and GPT-4 when answering question. For PL and FOL, the average reasoning chain length for ChatGPT at d_5 is 13.85, while for Gemini and GPT-4 at d_5 is 8.85 and 10.87, respectively. Longer reasoning chains do not necessarily correlate with better reasoning outcomes, highlighting the complexity of complex reasoning task. This suggests that optimizing reasoning chain length is crucial for improving model accuracy in complex scenarios.

[2] https://github.com/Mihir3009/Multi-LogiEval

Increasing Performance Trend in NM In our analysis of ChatGPT and the open-source model Yi-34B, we've observed consistent performance improvements with increasing depth in NM reasoning. This trend diverges from classical logic PL and FOL. Specifically, at depths d_2 to d_5 , NM exhibits novel performance due to unique rule combinations in reasoning patterns. For instance, at d_2, NM combines one PL rule with one NM reasoning pattern, progressing to two PL rules with one NM pattern at d_3, and so forth. The addition of NM reasoning patterns complements PL and FOL by providing supplementary evidence and improving contextual understanding. Notably, as depth increases, integrating basic classical rules with NM significantly enhances model accuracy, particularly evident at depths 4 and 5. This integration is pivotal for the notable performance gains observed in NM compared to classical logic at higher depths.

Larger Open-Source Models Show Decreased Performance Compared to Smaller Models Here, we examine Mistral-7B, Orca-13B, and Yi-34B, which differ significantly in parameter size. Mistral-7B, the smallest, performed best across various depths of classical logic, except at the simplest d_1. As reasoning depth increased, Mistral-7B consistently outperformed Orca-13B and Yi-34B, with Yi-34B only marginally better (1.5%) at d_3. For NM tasks, Mistral-7B and Yi-34B showed similar performance across all depths. At the most challenging depth (d_5) for both PL and FOL, Mistral-7B outperformed Orca-13B by achieving 3x performance despite Orca-13B's larger size. We believe that this capability of Mistral-7B is attributed to its architecture and training, enhancing its reasoning abilities, as discussed in Jiang et al. (2023). In particular, the training of Mistral-7B focused on enhancing its reasoning capabilities.

Lower Average Performance in FOL than PL at d_1 to d_5 Upon observing the reasoning chains with wrong final predictions for the FOL and PL, we find that the generic rules in FOL contexts lead to deviations from the correct reasoning path. In some cases, it assigns predicates incorrectly to the FOL inference rule. This pattern is more prominent at d_5, highlighting the large gap ($\sim 10\%$) in average performance between PL and FOL.

Lower Performance in d_1 of NM Reviewing the reasoning chains, we noted that models struggled to accurately map information. Interpreting various assumptions is crucial for effective reasoning at d_1. However, we observed that models have difficulty concluding based solely on assumptions present in the context when explicit knowledge is absent.

Preliminary Discussion on Multi-variable FOL Since our work focuses on evaluating LLMs' multi-step reasoning with simple FOL inference rules, we conducted only a preliminary study on their reasoning abilities for multi-variable FOL rules, discussed in Appendix H. This study reveals that creating natural language instances is challenging for this kind of setup.

Case Study on Evaluating *Multi-LogiEval* using Neural Symbolic Motivated by Olausson et al. (2023), we evaluate GPT-4 using the neuro-symbolic approach where we utilized the Prover9[3]. It first converts FOL statements into conjunctive normal form (CNF) and then performs resolution. Thus, we only evaluated data samples with FOL from *Multi-LogiEval*. For this study, we adapt the evaluation approach presented in Pan et al. (2023) where GPT-4 is used to convert the context and question in a formal executable program and use Prover9 to solve it. We used an implementation with a similar GPT-4 version compatible with Pan et al. (2023). We use the below prompt to convert natural language to an executable logic program:

> Given a problem description and a question. The task is to parse the problem and the question into first-order logic formulas. The grammar of the first-order logic formula is defined as follows :
>
> 1. logical conjunction: $expr1 \wedge expr2$
> 2. logical disjunction: $expr1 \vee expr2$
> 3. logical exclusive disjunction: $expr1 \oplus expr2$
> 4. logical negation: $\neg expr1$
> 5. $expr1$ implies $expr2$: $expr1 \rightarrow expr2$
> 6. $expr1$ if and only if $expr2$: $expr1 \leftrightarrow expr2$
> 7. logical universal quantification: $\forall x$
> 8. logical existential quantification: $\exists x$
>
> Output format: <logic form ::: description>

We evaluate the performance in terms of the accuracy of selecting the correct answer. We also report the "Executable Rate", which reflects the grammar correctness of the logical form, and the "Executable Accuracy" of the executable samples to measure the semantic correctness (Pan et al., 2023).

Reasoning Depth	Overall Acc (%)	Executable Rate (%)	Executable Acc (%)
d_1	55.83	85.83	51.46
d_2	46.67	76.67	47.83
d_3	38.89	77.78	30.00
d_4	40.00	77.78	41.43
d_5	60.00	77.78	62.86

Table 7: Performance Metrics by Reasoning Depths

As the reasoning depth increases from d_1 to d_4 (Table 7), there is a general trend of decreasing overall accuracy and executable accuracy, indicating that higher reasoning depth poses more challenges for executing code correctly. Interestingly, at d_5, both overall accuracy and executable accuracy show a significant improvement.

Human Evaluation and Further Discussion In this study, we performed a human evaluation on a selected subset of the *Multi-LogiEval*. Additionally, we explored potential strategies for enhancing LLMs' reasoning capabilities based on the findings of our current analysis. Please refer to Appendix I for further details on the human evaluation process and discussion.

5 Conclusions

In this work, we introduced *Multi-LogiEval*, a comprehensive multi-step logical reasoning benchmark consisting of three types of logic and over 60 combinations of inference rules. Our approach utilized two-stage methodology to construct data instances for our benchmark consisting of $\sim 1.6k$ data instances with $1 \sim 5$ reasoning depth. We evaluated a range of LLMs, including GPT-4, ChatGPT, Gemini, Yi, Orca, and Mistral on *Multi-LogiEval*. Experimental results revealed that these models struggle to perform logical reasoning, and their performance drops as the depth of logical reasoning increases (average accuracy of $\sim 68\%$ at d_1 to $\sim 43\%$ at d_5) for classical and non-classical logic. Furthermore, we systematically analyzed the reasoning chain generated by LLMs at various depths and presented interesting findings. We hope that *Multi-LogiEval* will facilitate future research in evaluating and enhancing the ability of existing and upcoming LLMs for multi-step logical reasoning.

[3] https://www.cs.unm.edu/~mccune/prover9/

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 five steps. *Multi-LogiEval* can be further extended by incorporating other inference rules and logic types, for instance, the inference rules in first-order logic that capture n-ary relations between multiple variables. We also note that this research is limited to the English language and can be extended to multilingual scenarios for evaluating the logical reasoning ability of LLMs.

Ethics Statement

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

Acknowledgement

We thank the anonymous reviewers for their constructive suggestions and feedback. We extend our gratitude to the Research Computing (RC), and Enterprise Technology at ASU for providing computing resources, and access to the ChatGPT enterprise version for experiments. We acknowledge support by a 2023 Spring Amazon Research Award (ARA). This material is also based upon work supported by the Engineering Research and Development Center - Information Technology Laboratory (ERDC-ITL) under Contract No. W912HZ24C0022.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3882–3890.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

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*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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*.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2023. Learning deductive reasoning from synthetic corpus based on formal logic. In *International Conference on Machine Learning*, pages 25254–25274. PMLR.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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.

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.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

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 Monotonic Logic Description

Propositional Logic (PL) 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 (FOL) FOL builds upon the foundations of PL 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 FOL as "There exists x such that x is John and x is a student." This logic extends the rules of PL, such as the *Modus Ponens* rule, which lets us infer conclusions for specific instances from general premises.

B Combinations of rules for Monotonic Logic

We created 27 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 8, 9, 10, and 11 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.

Generalized Rule Definition:
Rule 1: [if {P} is true then {Q} is true, and if {R} is true then {S} is true, and either {P} or {R} or both are true]
Rule 2: [if {S} is true, then {T} is true]

Formatting Instruction:
Complete the following tasks, only returning text in exactly the format given in the following examples.

Diversity Instruction:
Generate 5 more examples from multiple domains

Task Definition:
Task 1: Generate a short real life story that includes sentences to illustrate the above rules, replacing the entities P, Q, R, S, T with real values. Do not include the entity labels like P, Q, R, S, T in the story.
Task 2: Generate the following complex reasoning question using the story and the rules, by replacing the respective entities.
Q1: [If Q is not true, then is T true?]

Examples:

{P}: Mark takes Ryan to the park.
{Q}: Ryan gets to play on the swings.
{R}: Jackie makes cupcakes.
{S}: Jackie brings some cupcakes for Ryan.
{T}: Ryan gets a sugar rush.
Context: Mark promised to take his son Ryan to the park on Saturday. If Mark takes Ryan to the park, then Ryan will get to play on the swings. Jackie said she would make cupcakes on Saturday. If Jackie makes cupcakes, then she will bring some for Ryan. Ryan was certain that either Mark would take him to the park, or Jackie would make cupcakes, or both might be true. If Jackie ends up bringing cupcakes for Ryan, then Ryan will get a sugar rush.
Question: If Ryan did not get to play on the swings, then did Ryan get a sugar rush?

{P}: Jeff meditates regularly.
{Q}: Jeff improves his mental health.
{R}: Jeff eats healthy meals.
{S}: Jeff loses weight.
{T}: Jeff feels more confident about himself.
Context: Jeff wants to improve his health and fitness. If Jeff meditates regularly, he will improve his overall mental health. Also, if Jeff eats healthy nutritious meals, he is likely to lose weight. Jeff decides to either meditate regularly, or eat healthy meals, or do both simultaneously. He also knew that if he loses weight, then he will feel more confident about himself.
Question: If Jeff did not improve his mental health, did he feel more confident about himself?

{P}: Sam studies geography.
{Q}: Sam aces his geography test.
{R}: Sam travels the world.
{S}: Sam gains cultural awareness.
{T}: Sam becomes more open-minded.
Context: Sam wants to improve his geography knowledge. If he studies geography diligently, he will ace his geography test. Also, if Sam travels the world, he will gain cultural awareness from experiencing new places. Sam decides that he will either study geography, or travel the world, or do both. If Sam gains cultural awareness, he will become more open-minded.
Question: If Sam did not ace his geography test, did he become more open-minded?

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

C Example of Prompt

Figure 4 illustrates an example prompt for combination of rules from propositional logic, namely 'constructive dilemma' (CD), 'disjunctive syllogism' (DS), and 'modus ponens' (MP). CD is 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, and detailed 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 with their respective propositions, contexts, and questions.

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 8: 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:** ✗
DMT: $(\neg Q \lor \neg R) \dashv\vdash \neg(Q \land R)$ **CO:** $(P \to Q) \land (P \to R) \vdash P \to (Q \land R)$ **MT:** $(P \to (Q \land R) \land \neg(Q \land R) \vdash \neg P$	$(P \to Q), (P \to R)$	$\neg Q \lor \neg R$	**P:** ✗

Table 9: 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:** ✗
IM: $(P \to (Q \land R)) \vdash (P \land Q) \to R$ **MT:** $((P \land Q) \to R) \land \neg R \vdash \neg(P \land Q)$ **DMT:** $\neg(P \land Q) \vdash (\neg P \lor \neg Q)$ **DS:** $(\neg P \lor \neg Q) \land Q \vdash \neg P$	$(P \to (Q \land R))$	Q, $\neg R$	**P:** ✗

Table 10: 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 11: 5-step reasoning rule combinations for PL and FOL.

D 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 12 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 13 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.

E More Details on NM

Table 15 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 Priorities (RAP). These rules constitute our selection for depth-1 non-monotonic logical reasoning. Moving on to depths 2 through 5, we integrated classical and non-classical logic. Tables 16, 17, 18, and 19 outline the combinations of rules prepared respectively for depth-2, depth-3, depth-4, and depth-5 logical reasoning tasks. In this context, we combined BDR, DRD, PBD, DRO, REII, and REIII from non-monotonic logic with MP, MT, and DS from propositional logic to form combinations for depths 2 to 5 of data. Tables 20, 21, 22, and 23 show the prompts that we used to generate data instances respectively for depths 2, 3, 4, and 5. The instruction-based data generation can be seen in Tables 20, 21, 22, and 23. In addition to instruction-based generation, one-shot prompts were used for depth-3, depth-4, and depth-5 data generation as seen in Tables 21, 22, and 23.

F Validation of Data Instances

We involved four evaluators (who are also authors of this paper) for data validation. Each evaluator holds a graduate degree in computer science and has knowledge of logical reasoning. As discussed in Section 3.3, each sample is evaluated by one evaluator to ensure its logical correctness. We categorized errors into three distinct groups. The categories of errors identified are (i) Incorrect Logical Premises (ILP) which indicates that premises generated by the model in the context are logically incorrect (i.e., did not align with the intended conclusion), (ii) Leaking Conclusion (LC) where the context inadvertently revealed the conclusion, bypassing the need for the logical deduction, and (iii) Repetition of Samples (RS) where identical or nearly identical contexts are present, reducing dataset diversity. We found $\sim 14.3\%$ (223 samples) of the total 1552 samples with ILP, $\sim 3.7\%$ (57 samples) with LC, and $\sim 3.7\%$ (57 samples) with RS. We mitigated all these errors manually from the generated data instances to provide a high-quality evaluation set. Furthermore, we also analyzed the number of samples we corrected for PL ($\sim$ 22% - 115/525), FOL ($\sim$ 19% - 102/535), and NM ($\sim 25.9\%$ - 127/492), highlighting the difficulty of generating instances for specific logics. Similarly, we also analyzed depth-wise instance correction where we corrected $\sim 17.8\%$ (73/410), $\sim 21\%$ (93/442), $\sim 23.5\%$ (73/310), $\sim 33.2\%$ (93/280), and $\sim 21\%$ (23/110) for the depth d_1, d_2, d_3, d_4, and d_5, respectively, indicating the challenges of generating and validating multi-step reasoning context with increasing depth.

G Few shot evaluation Multi-LogiEval

We evaluate models in a few-shot setting (specifically, 3-shot) on Multi-LogiEval, revealing a notable enhancement in performance, as depicted in Table 24. In the 3-shot evaluation results, we observe notable improvements in the performance of various LLMs. GPT-4 consistently exhibits high accuracy across all depths, particularly excelling in PL and FOL. Though showing significant enhancements compared to its zero-shot performance across all the models, they still underperform in NM, highlighting a persistent challenge in this area. Open-source models such as Yi-34B and Mistral-7B, while benefiting from the 3-shot setup, still display noticeable performance drops in higher depths. Comparing these findings to the zero-shot results from Table 6, we see a general trend of improved performance in the 3-shot setting, indicating the effectiveness of few-shot prompting. However, the observed performance drop from d_4 to

Depth	Rules and Propositions	Context and Question
2	**Rules:** CD, DS **Propositions:** P: There is a big snowstorm coming. Q: Schools will be closed. R: Boss tells us to work from home. S: I avoid driving in the snow	**Context:** If there is a big snowstorm coming, schools will be closed tomorrow. Also, if my boss tells us to work from home, I can avoid driving in the snow. It seems either there will be a snowstorm or I'll be told to work from home, maybe both. **Question:** If schools were not closed tomorrow, then did I avoid driving in the snow?
3	**Rules:** BD, DS, MT **Propositions:** P: The weather is nice. Q: She goes for a walk. R: Finishes chores. S: Has free time. T: It's the weekend	**Context:** It was a beautiful sunny day. Amy knew that if the weather is nice, she goes for a walk. Amy also had chores to complete today. If Amy finishes her chores, then she has free time. Amy is certain that either the weather is nice, or she doesn't have free time, or the weather is nice and she doesn't have free time. She also knows that if it's the weekend, then she finishes her chores. **Question:** If Amy didn't go for a walk, then is it the weekend?
4	**Rules:** HS, CD, DS, MP **Propositions:** P: Studied hard for the exam. Q: Feel confident. R: Score well. S: Cooked nice dinner. T: Feel relaxed. U: Sleep soundly.	**Context:** Jim had a big exam coming up that he needed to prepare for. If Jim studied hard for the exam, he would feel confident going into it. If Jim felt confident about the exam, he would end up scoring well on it. His wife Lucy enjoyed cooking nice dinners. If Lucy cooked a nice dinner, she felt relaxed afterwards. Last night, either Jim studied hard, or Lucy cooked a nice dinner, or they both did those things. Jim knew that if Lucy felt relaxed after dinner, she always slept soundly through the night. **Question:** If Jim did not score well on the exam, did Lucy sleep soundly?
5	**Rules:** HS, MT, DS, MP, MP **Propositions:** P: Train consistently. Q: Increase endurance and stamina. R: complete the 26.2 mile marathon. S: Ate nutritious food. T: More steady energy. U: Train harder staying injury free.	**Context:** Jessica set a goal to run a marathon. She learned that if she trained consistently, she could increase her endurance and stamina. Jessica knew that if her endurance improved, she could complete the 26.2 mile marathon. To complement her training, Jessica made sure she either trained regularly, or ate nutritious foods, or did both. Eating nutritious foods gave Jessica more steady energy for her workouts. With this extra energy, Jessica found she could train harder while staying injury-free on her road to marathon success. **Question:** If Jessica does not complete the marathon, then does she stay injury-free during training?

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

d_5 in open-source models comparable across both settings, suggesting that while few-shot examples enhance overall accuracy, they do not fully mitigate the inherent challenges these models face in higher depths. Moreover, the performance trends identified in the zero-shot evaluation, such as the consistent decrease in accuracy for larger open-source models and the superior performance of proprietary models such as GPT-4 and ChatGPT in PL and FOL, remain similar in the 3-shot setting.

H Extended first-order logic with n-ary relations

First-order logic often involves handling n-ary relations involving more than two variables—such as the ternary relation in "If $P(a, b, c) \wedge Q(c, d)$ then $R(a, d)$". Moreover, one can alternate *for all* ($\forall$), and *there exists* ($\exists$) for any number of times in FOL, and that means there are an infinite number of such rules in first-order logic. As discussed in section 3.1, our aim is not to build a comprehensive set covering all the possible inference rules but rather to evaluate the reasoning ability of language models up to a reasoning depth of five on a systematically curated set of inference rules. However, to evaluate the ability of LLMs to reason with such complex rules, we explore 7 such inference rules for which we generated data using a similar prompt structure as depicted in Figure 3. We generate 10 instances for each of the inference rule, resulting in 70 instances for evaluation. The choice of inference rules can be found in Table 14. We evaluate the large-scale models GPT-4, ChatGPT, and Gemini. These models achieve an average accuracy of 80%, 84.3%, and 90%, respectively. This demonstrates that these LLMs can comprehend multi-variable FOL, but the rules currently involve only single-step reasoning. Our work also shows that these models perform well with single-step reasoning. Exploring multi-step reasoning with multi-variable FOL presents an interesting direction for future research direction.

Depth	Rules and Predicates	Context and Question
2	**Rules:** CD, DS **Predicates:** P: Compose original music. Q: Work would be Unique. R: Promote music online. S: Gain following.	**Context:** An aspiring musician decided to try writing their own songs. They realized that if they composed original music, their work would be unique; if they promoted their music online, they would gain a following. The musician could write original songs or promote their music online. **Question:** Given that Maria's music was not unique, is it true that she gained a following online?
3	**Rules:** BD, DS, MT **Predicates:** P: It's Monday. Q: There is a staff meeting. R: Finish report. S: Submit the report. T: Good Employee.	**Context:** It was a busy morning at the office. If it was Monday, then there would be a staff meeting. If they finished the report, then they could submit it to their manager. They were certain that either it was Monday, or they did not submit the report. It is known at the office that if someone is a good employee, they finish their reports on time. **Question:** Sam did not have a staff meeting, is Sam a good employee?
4	**Rules:** BD, DS, MT, DS **Predicates:** P: First day of school. Q: students feel nervous and excited. R: Study Hard. S: get good grades. T: teacher is very strict. U: class textbook is very long.	**Context:** If it is the first day of school, then students feel nervous and excited. If someone studies hard, then they get good grades. Either it is the first day, or they do not get good grades, or it is the first day and they do not get good grades. If a teacher is very strict, then students have to study hard for that class. Either the teacher is very strict, or the class textbook is very long, or perhaps both are true. **Question:** Emma was not nervous on the first day, does this mean did she have a very long textbook in one of her classes?
5	**Rules:** HS, MT, DS, MP, MP **Predicates:** P: Practice drawing techniques. Q: improve artistic skills. R: sell their artwork. S: studies art history and famous artists. T: gain inspiration. U: develop creative style.	**Context:** Someone wanted to become an artist. They learned that if they practiced drawing techniques consistently, they would improve their artistic skills. With improved artistic skills, they could sell their artworks. Either someone practices drawing techniques consistently, or someone studies art history and famous artists, or they do both. If someone studies art history and famous artists, then they gain inspiration for their own art. If they gain inspiration, then they can develop their own creative style. **Question:** If Emma cannot sell her artworks yet, then has she developed her own creative style?

Table 13: Natural language examples of rule combinations of each depth for FOL.

I Human Evaluation and Discussion

We have conducted a human evaluation on a subset of Multi-LogiEval. Specifically, we selected 15 unique instances covering all 5 depths (5 instances for each logic type) from Multi-LogiEval. This selection resulted in a total instances of 75 <context, question> pairs. We hired three graduate student volunteers to provide the evaluations. Task instructions provided to all three annotators are similar to prompts provided to LLMs. Each instance pair is answered/annotated by three different annotators with 0.853 inter-annotator agreement (measured with raw/observed agreement). Here are the results for three logic types averaged across three annotators for each depth. The average accuracies are d_1- 75.56, d_2 - 64.71, d_3 - 64.44, d_4 - 66.67, and d_5 - 64.44. From the results, we can observe that humans perform better at d_1 compared to higher depths. For higher depths, the human performance is low and consistent.

Discussion on Future Work Our results provide deeper insights into LLMs' logical reasoning abilities by analyzing reasoning chains (Section 4.3) manually to some extent. Specifically, we analyzed where these models make mistakes and what are their limitations. We believe that such insights can help design better pre-training or alignment strategies to improve the reasoning abilities of LLMs. For instance, during pre-training, logical connectives can be treated differently at various stages of the transformer architecture. Additionally, in alignment techniques involving preference optimization, preference data can be created to prefer outputs with more logical correctness. The proposed approaches and findings in our paper can help in creating such datasets. Furthermore, exploring newer techniques such as DPO (Rafailov et al., 2024), and KTO (Ethayarajh et al., 2024) for their suitability in improving logical reasoning also can be an interesting future direction.

Rule	Extended First-order Logic with Multi-variable
1	$\forall x \forall y((p(x) \land q(x)) \to r(x,y)) \land \exists u \exists v(p(u) \land \neg r(u,v)) \vdash \exists y \neg q(y)$
2	$\forall x \forall y((p(x) \land q(x)) \to \neg s(x,y)) \land \forall z(r(z) \to p(z)) \land r(a) \land s(a,b) \vdash \neg q(b)$
3	$\forall x \exists y((p(x) \to q(x,y)) \land \forall u \forall v((q(u,v) \land r(u,v)) \to s(v)) \land \exists z \exists k(p(z) \land r(z,k)) \vdash \exists w s(w)$
4	$\forall x \forall y \forall z(p(x,y,z) \to (q(x,z) \lor r(y))) \land \exists u \exists v \exists w(p(u,v,w) \land \neg q(u,w)) \vdash \exists s r(s)$
5	$\forall x((p(x) \to \exists y r(y,x)) \land p(a) \vdash \exists z r(z,a)$
6	$\forall x \forall y(p(x,y) \lor q(x,y)) \land \exists u \exists v \neg q(u,v) \vdash \exists z \exists w p(z,w)$
7	$\forall x \forall y(p(x,y) \to (q(x) \land r(y)) \land p(a,b) \vdash q(a) \land r(b))$

Table 14: FOL inference rules that establish the relationship between multiple variables

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

Rule	Examples
BDR_MP Conclusion of BDR: X MP: $(X \rightarrow Y) \wedge X \vdash Y$	**Context:** Jim and Pam work at the same office. Normally, employees at that office get free lunch. Jim does not get free lunch. If Pam gets free lunch, then she gets an hour lunch break. **Question:** Can we conclude Pam gets an hour lunch break? (Yes)
BDR_MT Conclusion of BDR: X MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	**Context:** Emma and Jacob are students in the same class. Usually students in that class submit homework assignments. Emma did not submit the last homework. If Jacob missed over 3 classes, that means he likely did not submit the homework. **Question:** Can we conclude Jacob missed over 3 classes? (No)
DRD_MP Conclusion of DRD: X MP: $(X \rightarrow Y) \wedge X \vdash Y$	**Context:** The Honda and Toyota are sedans. Sedans normally have four doors. The Honda might not have four doors even though it's a sedan. If the Toyota has a four doors then it has four windows. **Question:** Can we conclude the Toyota likely has four windows? (Yes)
DRD_MT Conclusion of DRD: X MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	**Context:** Oaks and pines are types of trees. Typically trees grow from seeds. Oaks may not grow from seeds even though they are trees. If a pine is artificial, then it does not grow from a seed. **Question:** Can we conclude the pine is artificial? (No)
DRI_MP Conclusion of DRI: X MP: $(X \rightarrow Y) \wedge X \vdash Y$	**Context:** John and Mary are students in the same class. Usually students in their class do homework every day. John did not do his homework yesterday. Mary studied extra material last night. If Mary did her usual homework, she would have also reviewed her notes. **Question:** Can we conclude that Mary reviewed her notes last night? (Yes)
DRI_MT Conclusion of DRI: X MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	**Context:** Sara and David ordered dessert at a restaurant. Usually, people who order dessert also order coffee. Sara did not order coffee. David requested extra whipped cream. If a customer asks for extra toppings, it means they did not order coffee. **Question:** Can we conclude David asked for extra toppings? (No)
PBD_MP Conclusion of PBD: X MP: $(X \rightarrow Y) \wedge X \vdash Y$	**Context:** Jenny said the dog dug up the flower bed. Her brother said the dog did not dig up the flower bed. People usually tell the truth. Jenny is more trustworthy than her brother. If the dog dug up the flowers, it likely made a mess. **Question:** Can we conclude the dog made a mess? (Yes)
PBD_MT Conclusion of PBD: X MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	**Context:** John said the shirt was blue. Mary said the shirt was not blue. Normally people are correct when they make assertions. John had a closer look at the shirt than Mary. If the shirt was purple, it could not be blue. **Question:** Can we conclude the shirt was purple? (No)
REI_MP Conclusion of REI: X MP: $(X \rightarrow Y) \wedge X \vdash Y$	**Context:** Ben, Mark, and Jacob took a history test. Students who study many hours usually pass history tests. Ben and Mark did not study many hours. If Jacob passed the history test, he must have paid attention in class. **Question:** Can we conclude Jacob paid attention in class? (Yes)
REI_MT Conclusion of REI: X MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	**Context:** John, Peter and Kate are students in math class. Students in math class normally do homework. John and Peter did not do their math homework. If Kate missed class then she did not do her math homework. **Question:** Can we conclude Kate missed class? (No)
REII_MP Conclusion of REII: X MP: $(X \rightarrow Y) \wedge X \vdash Y$	**Context:** John bought a new phone. New phones usually come with a warranty. However, some new phones do not come with a warranty. If a phone has a warranty, then it has customer support. **Question:** Can we conclude John's new phone has customer support? (Yes)
REII_MT Conclusion of REII: X MT: $(X \rightarrow Y) \wedge \neg Y \vdash \neg X$	**Context:** Kate booked a room at hotel Y. Rooms at hotel Y are usually clean. There is at least one room at hotel Y that is not clean. If Kate's room has mold, then it is probably not clean. **Question:** Can we conclude Kate's room has mold? (No)

Table 16: Natural language examples of rule combinations of depth-2 for NM.

Rule	Examples
Rule: d3_1 **Assumptions:** 1: A and B are objects of type T and have property S. 2: Normally objects of type T with property S have property U. 3: if A has property U implies C has property D 4: if C has property D implies E has property F **Question 1:** Can we conclude if E does not have F then B has U? (YES) **Question 2:** Can we conclude if E does not have F then B does not have U? (NO)	**Context:** Smartphone A and Smartphone B both have GPS technology. Normally, smartphones with GPS technology also have internet connectivity. If smartphone A has internet connectivity, then Mike can access online maps. If Mike can access online maps, then Emily can get driving directions from Mike. **Question 1:** Can we conclude if Emily can not get driving directions from Mike, then smartphone B has internet connectivity? (Yes) **Question 2:** Can we conclude if Emily can not get driving directions from Mike, then smartphone B does not have internet connectivity? (No)
Rule: d3_2 **Assumptions:** 1: A and B are objects of type T and have property S. 2: Normally objects of type T with property S have property U. 3: if C has property G implies C has property D 4: if A has property U implies E has property F 5: either C has property G or E does not have property F or both **Question 1:** Can we conclude if C does not have D then B has U? (YES) **Question 2:** Can we conclude if C does not have D then B does not have U? (NO)	**Context:** Car A and car B are electric vehicles. Normally, electric vehicles (cars) have fast-charging capabilities. If car C is a hybrid, then car C has good fuel efficiency. If car A has a fast-charging capability, then it implies that the environment is very eco-friendly. Either car C is a hybrid or the environment is not very eco-friendly, or both. **Question 1:** Can we conclude if car C is not a good fuel efficient then Car B has a fast-charging capability? (Yes) **Question 2:** Can we conclude if car C is not a good fuel efficient then Car B does not have a fast-charging capability? (No)

Table 17: Natural language examples of rule combinations of depth-3 for NM.

Rule	Examples
Rule: d4_1 **Assumptions:** 1: A and B are objects of type T and have property S. 2: Normally objects of type T with property S have property U. 3: if C has property G implies C has property D 4: if E has property L implies E has property F 5: either C has property G or E has property F or both 6: if A has property U then E has property L **Question 1:** Can we conclude if C does not have D then B has U? (YES) **Question 2:** Can we conclude if C does not have D then B does not have U? (NO)	**Context:** Apple tree and Orange tree are fruit trees. Normally, fruit trees produce edible fruit. If Garden is regularly watered, then its plants are flourishing. If Orchard receives enough sunlight, then it yields high-quality fruit. Either Garden has regular watering or Orchard yields high-quality fruit or both. If the apple tree produces edible fruit, then Orchard receives enough sunlight. **Question 1:** Can we conclude if Garden does not have flourishing plants then the orange tree produces edible fruit? (Yes) **Question 2:** Can we conclude if Garden does not have flourishing plants then the orange tree does not produce edible fruit? (No)
Rule: d4_2 **Assumptions:** 1: A and B are objects of type T and have property S. 2: Normally objects of type T with property S have property U. 3: if C has property G implies C has property D 4: if E has property L implies E has property F 5: either C does not have property D or E does not have property F or both 6: if A has property U then E has property L **Question 1:** Can we conclude if C has property G then B has U? (YES) **Question 2:** Can we conclude if C does not have G then B does not have U? (NO)	**Context:** Assume A and B are plants of species T and they both produce flowers. Normally, flowering plants of species T also bear fruit. If an animal C is a bird, then it can fly. If an environment has a lot of sunlight, then it supports plant growth. Either the bird cannot fly or the environment does not support plant growth or both. If plant A bears fruit, then the environment has a lot of sunlight. **Question 1:** Can we conclude if the bird is capable of flying then plant B bears fruit? (Yes) **Question 2:** Can we conclude if the bird can fly then plant B does not bear fruit? (No)

Table 18: Natural language examples of rule combinations of depth-4 for NM.

Rule	Examples
Rule: d5_1 **Assumptions:** 1: A and B are objects of type T and have property S. 2: Normally objects of type T with property S have property U. 3: if C has property G implies C has property D 4: if E has property L implies E has property F 5: either C has property G or E has property F or both 6: if I has property H then E has property L 7: if A has property U then I has property H **Question 1:** Can we conclude if C does not have D then B has U? (YES) **Question 2:** Can we conclude if C does not have D then B does not have U? (NO)	**Context:** Rose and Lily are plants that flower. Normally, plants that flower also produce seeds. If a plant is a Cactus, and it has thorns, then it can survive in the desert. If a plant is an Orchid, and it has broad leaves, then it can grow in tropical areas. Either a Cactus has thorns, or an Orchid can grow in tropical areas, or both. If a Lotus has flowers, then an Orchid has broad leaves. If a Rose produces seeds, then a Lotus has flowers. **Question 1:** Can we conclude if a Cactus cannot survive in the desert then a Lily produces seeds? (YES) **Question 2:** Can we conclude if a Cactus cannot survive in the desert then a Lily does not produce seeds? (NO)

Table 19: Natural language examples of rule combinations of depth-5 for NM.

Rule:
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.
4: If B has property Q then it implies B has property C.
Question: Can we conclude B has property C?

Task 1:
Generate a short generic story that should only contain the natural language sentences for assumptions 1, 2, 3, and 4 using propositions to replace the labels A, B and so on.
The story should not include labels like p or q and so on.

Task 2:
Generate the question by replacing them with the entities with respective propositions.

Table 20: An example of prompt used to generate data instance for depth-2 using NM-BDR and PL-MP

Rule: d3_1

Assumptions:
1: A and B are objects of type T and have property S.
2: Normally objects of type T with property S have property U.
3: if A has property U implies C has property D
4: if C has property D implies E has property F

Question 1:
Can we conclude if E does not have F then B has U? (YES)

Question 2:
Can we conclude if E does not have F then B does not have U? (NO)

Task 1: Generate a short context paragraph by replacing all the entity labels A, B, and so on in the above context with propositions and real entities. The generated context should have natural language sentences for all the sentences 1-4. It should not include label representations like A or B and should not mention the words "property".

Task 2: Generate questions 1 and 2 by replacing the respective labels from the generated context.

Example 1:
Assumptions:
Smartphone A and Smartphone B both have GPS technology.
Normally, smartphones with GPS technology also have internet connectivity.
If smartphone A has internet connectivity, then Mike can access online maps.
If Mike can access online maps, then Emily can get driving directions from Mike.

Question 1:
Can we conclude if Emily can not get driving directions from Mike, then smartphone B has internet connectivity?

Question 2:
Can we conclude if Emily can not get driving directions from Mike, then smartphone B does not have internet connectivity?

Table 21: An example of prompt used to generate data instance for depth-3 for NM

Rule: d4_1

Assumptions:
1: A and B are objects of type T and have property S.
2: Normally objects of type T with property S have property U.
3: if C has property G implies C has property D
4: if E has property L implies E has property F
5: either C has property G or E has property F or both
6: if A has property U then E has property L

Question 1:
Can we conclude if C does not have D then B has U? (YES)

Question 2:
Can we conclude if C does not have D then B does not have U? (NO)

Task 1: Generate a short context paragraph by replacing all the entity labels A, B, and so on in the above context with propositions and real entities. The generated context should have natural language sentences for all the sentences 1-4. It should not include label representations like A or B and should not mention the words "property".

Task 2: Generate questions 1 and 2 by replacing the respective labels from the generated context.

Example 1:
Assumptions:
Apple tree and Orange tree are fruit trees.
Normally, fruit trees produce edible fruit.
If Garden is regularly watered, then its plants are flourishing.
If Orchard receives enough sunlight, then it yields high-quality fruit.
Either Garden has regular watering or Orchard yields high-quality fruit or both.
If the apple tree produces edible fruit, then Orchard receives enough sunlight.

Question 1: Can we conclude if Garden does not have flourishing plants then the orange tree produces edible fruit?

Question 2: Can we conclude if Garden does not have flourishing plants then the orange tree does not produce edible fruit?

Table 22: An example of prompt used to generate data instance for depth-4 for NM

Rule: d5_1

Assumptions:
1: A and B are objects of type T and have property S.
2: Normally objects of type T with property S have property U.
3: if C has property G implies C has property D
4: if E has property L implies E has property F
5: either C has property G or E has property F or both
6: if I has property H then E has property L
7: if A has property U then I has property H

Question 1:
Can we conclude if C does not have D then B has U? (YES)

Question 2:
Can we conclude if C does not have D then B does not have U? (NO)

Task 1: Generate a short context paragraph by replacing all the entity labels A, B, and so on in the above context with propositions and real entities. The generated context should have natural language sentences for all the sentences 1-4. It should not include label representations like A or B and should not mention the words "property".

Task 2: Generate questions 1 and 2 by replacing the respective labels from the generated context.

Example 1:
Assumptions:
Rose and Lily are plants that flower.
Normally, plants that flower also produce seeds.
If a plant is a Cactus, and it has thorns, then it can survive in the desert.
If a plant is an Orchid, and it has broad leaves, then it can grow in tropical areas.
Either a Cactus has thorns, or an Orchid can grow in tropical areas, or both.
If a Lotus has flowers, then an Orchid has broad leaves.
If a Rose produces seeds, then a Lotus has flowers.

Question 1: Can we conclude if a Cactus cannot survive in the desert then a Lily produces seeds? (YES)

Question 2: Can we conclude if a Cactus cannot survive in the desert then a Lily does not produce seeds? (NO)

Table 23: An example of prompt used to generate data instance for depth-5 for NM

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	d_3	d_4	d_5
GPT-4	90.00	85.71	84.44	79.17	73.33	97.78	84.76	73.33	68.33	73.33	54.38	56.11	75.00	90.00	75.00
ChatGPT	96.67	82.86	77.78	79.17	80.00	94.44	86.67	84.44	64.17	64.44	45.63	41.67	57.50	65.00	45.00
Gemini	92.22	73.33	81.48	88.33	77.78	90.00	83.81	81.48	76.67	57.78	59.38	42.78	75.00	62.50	75.00
Yi-34B	68.89	61.90	66.67	64.17	64.44	76.67	61.90	62.96	45.00	51.11	59.38	33.33	52.50	52.50	50.00
Orca-13B	85.56	80.00	72.59	75.83	68.89	91.11	73.33	63.70	55.00	42.22	56.88	46.67	60.00	50.00	50.00
Mistral-7B	80.00	64.76	71.11	73.33	66.67	93.33	71.43	62.96	62.50	42.22	37.50	36.11	45.00	57.50	70.00
Avg	**84.22**	**75.05**	**74.52**	**74.33**	**70.67**	**90.67**	**75.62**	**69.48**	**59.00**	**54.66**	**50.75**	**42.78**	**60.83**	**62.92**	**60.83**

Table 24: Few-shot Evaluation of LLMs in terms of accuracy on *Multi-LogiEval*.