Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models

Anonymous ACL submission

Abstract

Recently developed large language models (LLMs) have been shown to perform remarkably well on a wide range of language understanding tasks. But, can they really "reason" over the natural language? This question has been receiving significant research attention and many reasoning skills such as commonsense, numerical, and qualitative have been studied. However, the crucial skill pertaining to 'logical reasoning' has remained underexplored. Existing work investigating this reasoning ability of LLMs has focused only on a couple of inference rules (such as modus ponens and modus tollens) of propositional and first-order logic. Addressing the above limitation, we comprehensively evaluate the logical reasoning ability of LLMs on 25 different reasoning patterns spanning over propositional, first-order, and non-monotonic logics. To enable systematic evaluation, we introduce *LogicBench*, a natural language question-answering dataset focusing on the use of a single inference rule. We conduct detailed analysis with a range of LLMs such as GPT-4, ChatGPT, Gemini, Llama-2, and Mistral using chain-of-thought prompting. Experimental results show that existing LLMs do not fare well on *LogicBench*; especially, they struggle with instances involving complex reasoning and negations. Furthermore, they sometimes tend to prioritize parametric knowledge over contextual information and overlook the correct reasoning chain. We believe that our work and findings facilitate future research for evaluating and enhancing the logical reasoning ability of LLMs[1].

1 Introduction

Large language models such as GPT-4, ChatGPT, Google Gemini, Llama-2 (Touvron et al., 2023), and Mistral (Jiang et al., 2023) have made remarkable progress in NLP research enabling machines to perform 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 reason "logically" over natural language text remains under-explored, even though logical reasoning is a fundamental aspect of intelligence and a crucial requirement for many practical applications, such as question-answering systems (Khashabi, 2019) and conversational agents (Beygi et al., 2022). Although several datasets have been proposed (Clark et al., 2021; Tian et al., 2021; Joshi et al., 2020; Saeed et al., 2021) to evaluate the logical reasoning capabilities of LLMs, these datasets are limited in their scope by (1) not evaluating logical reasoning independently of other forms of reasoning such as LogiQA (Liu et al., 2021a) and ReClor (Yu et al., 2020); and (2) evaluating only a single type of logic and covering only few logical inference rules as done in FOLIO (Han et al., 2022) and ProntoQA (Saparov and He, 2023). Thus, our aim in this work is to address the lacuna of having a more comprehensive set of inference rules for evaluating the logical reasoning ability of LLMs.

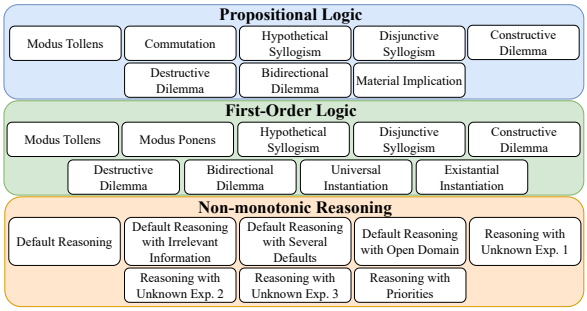

Figure 1: Comprehensive representation of different inference rules and reasoning patterns covered by propositional, first-order, and non-monotonic logics. *Exp.* indicates Expectation

To this end, we introduce a systematically created question-answering dataset for the evaluation of logical reasoning ability using a single inference

[1]Data and code are available at https://anonymous.4open.science/r/LogicBench-EEBB

rule, called *LogicBench*. Besides evaluating the logical reasoning ability of LLMs, by evaluating models on single inference rules, we can also gain insights into the frequency of text sequences corresponding to these rules in the pre-training data and their impact on model performance. As illustrated in Figure 1, *LogicBench* includes a total of 25 reasoning patterns across 'propositional, first-order, and non-monotonic' logics. To the best of the authors' knowledge, this is the first work to study non-monotonic reasoning, as well as various inference rules in propositional and first-order logics including hypothetical and disjunctive syllogism; and bidirectional, constructive, and destructive dilemmas in NLP domain. To evaluate LLMs using *LogicBench*, we formulate two different tasks: (i) a Binary Question-Answering (BQA) task in which the context comprises logical statements and the models have to determine whether a conclusion given in the question is logically entailed by the context, and (ii) a Multiple-Choice Questions-Answering (MCQA) task where models must select the most appropriate logical conclusion from four distinct options, based on the provided context. Examples instances of various reasoning patterns are presented in Table 4 and App. C.

To construct *LogicBench*, we use a three-stage procedure (refer to §3). In the first stage, we prompt GPT-3.5 to generate a variety of coherent natural language sentences having different 'ontologies' (i.e., a collection of concepts as car, person, and animals) and their corresponding negations (refer to §3.2.1). In the second stage, we generate *(context, question)* pairs where the context represents a natural language narrative consisting of logical statements, and the question is formulated to exhibit the logical conclusion derived from the context. In the third stage, we generate task-specific data instances (i.e., *(context, question, answer)* triplets).

We conduct a comprehensive evaluation with a range of LLMs on *LogicBench* including GPT-4, ChatGPT (GPT-3.5-Turbo), Gemini-Pro, Llama-2-7B-Chat, and Mistral-7B-Instruct using chain-of-thought (Wei et al., 2022). In particular, we measure the accuracy of LLMs predictions on both BQA and MCQA tasks. Our experiments result in several interesting findings such as LLMs often struggle to reason over complex logical contexts and encounter difficulties with inference rules involving negations. Experimental results reveal that these models struggle with respect to many of the inference rules and patterns, suggesting significant room for improvement in their logical reasoning abilities. To further demonstrate the use of *LogicBench*, we synthetically augment it and fine-tune T5-large. Our preliminary results (App. I) show that this improves the logical reasoning ability of existing models leading to performance improvement on other logic datasets, LogicNLI, and FOLIO ($\sim 2\%$ on an average), and shows competitive performance on LogiQA and ReClor.

Dataset	Logic Covered			Inference Rules/Axioms Provided with Data	Generation Code Available
	PL	FL	NM		
Ruletaker	✗	✓	✗	✗	Human-annotated
LogicNLI	✗	✓	✗	✗	Semi-automated
ProofWriter	✓	✓	✗	✗	✗
FOLIO	✗	✓	✗	✗	Human-annotated
SimpleLogic	✓	✗	✗	✗	✓
ProntoQA	✗	✓	✗	✓	✓
LogicBench	✓	✓	✓	✓	✓

Table 1: *LogicBench*'s comparison with current datasets

2 Related Work

As LLMs continue to evolve rapidly, it becomes increasingly crucial to evaluate their diverse reasoning capabilities, as well as those of forthcoming LLMs. LogiQA (Liu et al., 2021a) and ReClor (Yu et al., 2020) have made notable contributions by compiling multichoice questions from standardized examinations that demand diverse forms of logical reasoning. In contrast to *LogicBench*, these datasets involve mixed forms of reasoning and do not focus on assessing logical reasoning in isolation.

A few past attempts have been made to evaluate only logical reasoning while excluding other forms of reasoning. For example, CLUTTER (Sinha et al., 2019) covers inductive reasoning, (Hahn et al., 2021) covers temporal logic, and Ruletaker (Clark et al., 2021) evaluates whether a transformer-based model emulates deductive reasoning over synthetically generated statements in a limited setting. LogicNLI (Tian et al., 2021) introduced a diagnostic benchmark for FOL reasoning, with the dataset constructed by automatically generating logic expressions and replacing the entity and attribute placeholders.

Our proposed dataset is similar (in terms of task formulation) to ProofWriter (Tafjord et al., 2021), FOLIO (Han et al., 2022), and ProntoQA (Saparov and He, 2023) which are QA datasets designed to test reasoning ability. ProofWriter provides multi-hop proofs for each example, while FOLIO gives diverse and complex logical expressions, however, it is only limited to FOL. ProntoQA (Saparov and

Names	Propositional Logic	Extension to a (restricted) First-order Logic
MP	$((p \rightarrow q) \wedge p) \vdash q$	$(\forall x(p(x) \rightarrow q(x)) \wedge p(a)) \vdash q(a)$
MT	$((p \rightarrow q) \wedge \neg q) \vdash \neg p$	$(\forall x(p(x) \rightarrow q(x)) \wedge \neg q(a)) \vdash \neg p(a)$
HS	$((p \rightarrow q)) \wedge (q \rightarrow r)) \vdash (p \rightarrow r)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (q(x) \rightarrow r(x))) \vdash (p(a) \rightarrow r(a))$
DS	$((p \vee q) \wedge \neg p) \vdash q$	$(\forall x(p(x) \vee q(x)) \wedge \neg p(a)) \vdash q(a)$
CD	$((p \rightarrow q) \wedge (r \rightarrow s) \wedge (p \vee r)) \vdash (q \vee s)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (r(x) \rightarrow s(x))) \wedge (p(a) \vee r(a))) \vdash (q(a) \vee s(a))$
DD	$((p \rightarrow q) \wedge (r \rightarrow s) \wedge (\neg q \vee \neg s)) \vdash (\neg p \vee \neg r)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (r(x) \rightarrow s(x))) \wedge (\neg q(a) \vee \neg s(a))) \vdash (\neg p(a) \vee \neg r(a))$
BD	$((p \rightarrow q) \wedge (r \rightarrow s) \wedge (p \vee \neg s)) \vdash (q \vee \neg r)$	$(\forall x((p(x) \rightarrow q(x)) \wedge (r(x) \rightarrow s(x))) \wedge (p(a) \vee \neg s(a))) \vdash (q(a) \vee \neg r(a))$
CT	$(p \vee q) \vdash (q \vee p)$	-
MI	$(p \rightarrow q) \vdash (\neg p \vee q)$	-
EG	-	$P(a) \Rightarrow \exists x P(x)$
UI	-	$\forall x\, A \Rightarrow A\{x \mapsto a\}$

Table 2: Inference rules and (two) axioms that establish the relationship between premises and conclusions. MP: Modus Ponens, MT: Modus Tollens, HS: Hypothetical Syllogism, DS: Disjunctive Syllogism, CD: Constructive Dilemma, DD: Destructive Dilemma, BD: Bidirectional Dilemma, CT: Commutation, MI: Material Implication, EG: Existential Generalization, UI: Universal Instantiation

He, 2023) provides explanation and reasoning steps but is limited to modus ponens in FOL. Nevertheless, several crucial attributes motivated us to create *LogicBench* (see Table 1 for comparison). Additional datasets for evaluating logical reasoning also exist such as SimpleLogic (Zhang et al., 2022) provides a class of logical reasoning problems, TaxiNLI (Joshi et al., 2020) introduces logical taxonomy in NLI task and RuleBert (Saeed et al., 2021) covers only soft logical rules. In summary, *LogicBench* evaluates logical reasoning in isolation and provides diverse inference rules and logic types compared to existing datasets. Extended related work is discussed in App. B.

3 LogicBench

3.1 Logics Types

Propositional Logic (PL) Propositional logic employs a collection of statements or propositions (denoted as $\mathcal{P} = p_1, p_2, ..., p_n$, where p_i represents a proposition) and builds upon them using logical connectives such as '$\wedge$', '$\vee$', '$\rightarrow$', '$\leftrightarrow$', and '$\neg$'. Several inference rules for propositional logic have been defined using which given a set of premises, one can derive a sound conclusion. For example, let's consider two propositions: p_1, which states "It is raining," and p_2, which states "It is cloudy." From these propositions, we can construct a context/knowledge base (KB) with two premises: (1) $p_1 \rightarrow p_2$ and (2) p_1. With this KB, we can conclude p_2. This inference rule is written as $((p_1 \rightarrow p_2) \wedge p_1) \vdash p_2$ and is known as 'Modus Ponens'. In our study, we explore nine distinct inference rules of propositional logic, extensions of seven of them with one-variable and a universal quantifier, and two axioms of first-order logic as shown in Table 2. These inference rules provide a proper framework for deriving valid conclusions.

First-order Logic (FOL) In this work, we consider a restricted set of logical axioms for FOL that utilize quantifiers, $\forall$ (universal quantifier) and $\exists$ (existential quantifier). The universal quantifier ($\forall$) denotes that a statement holds true for all instances within a specific category. In contrast, the existential quantifier ($\exists$) indicates that a statement is true for at least one instance within its scope. For instance, a simple extension of propositional 'Modus Ponens' is an inference rule where given the premises $\forall(p(x) \rightarrow q(x))$ and $p(a)$, we conclude $q(a)$ (e.g., given "All kings are greedy" and "Sam is a king", we can conclude "Sam is greedy"). Here, we explore two axioms (EG and UI - in detail in App. C.3) and various inference rules that incorporate the quantifiers (shown in Table 2).

Non-monotonic (NM) Reasoning In this work, we analyze a range of logical reasoning templates in NM logics involving "Default Reasoning," "Reasoning about Unknown Expectations," and "Reasoning about Priorities." These templates are inspired by the compilation (Lifschitz, 1989) made in 1989 to evaluate the abilities of various non-monotonic logics that were being developed at that time. Below Table 3 shows examples of NM reasoning. Additional examples are given in App. C.4.

A key aspect of NM logics is to formalize notions such as "normally," "typically," and "usually" that are not directly formalizable using classical quantifiers in the first-order setting. The general rule "Heavy blocks are normally located on the table" does not imply that "All heavy blocks are al-

Basic Default Reasoning	Default Reasoning with Irrelevant Information
Context: Block A and block B are both heavy objects that are typically found on the table. However, there is a possibility that block A might not follow this usual convention. It is important to note this exception. Conclusion: B is on the table.	Context: In a room filled with various objects, two heavy blocks, block A and block B, stand out. Normally, heavy blocks like these are placed on the table, but surprisingly, block A is not found on the table. On the other hand, block B grabs attention with its vibrant red color. Conclusion: B is on the table.
Reasoning about Unknown Expectations	**Reasoning about Priorities**
Context: In this situation, there are three heavy blocks: A, B, and C. Typically, heavy blocks are found on the table. However, it is known that at least one of the blocks, either A or B, is not currently on the table. Conclusion: C is on the table. Exactly one of A, B is not on the table.	Context: John confidently states that the vehicle is situated in the driveway, while Sara adamantly counters, asserting that it is not parked inside the garage. Conclusion: If John's evidence is more reliable than Sara's then the car is parked in the driveway.

Table 3: Illustrative examples of non-monotonic reasoning adapted from (Lifschitz, 1989).

ways located on the table". Rather, this rule allows for exceptions. Our work explores various NM reasoning patterns, as depicted in Figure 1, to delve deeper into the nuances of this type of reasoning.

3.2 Data Creation

Our data creation procedure, illustrated in Figure 2, consists of three stages:

1. Sentence Generation: Starting with a given prompt, we generate coherent sentences and their negations that incorporate different ontologies.

2. NL Conversion: Pairs of *(context, question)* are generated using pre-defined templates from which context is then converted to a natural language narrative using the prompt.

3. Task Instance Generation: Task-specific *(context, question, answer)* triplets are generated. BQA requires answers in the form of "yes" or "no", whereas MCQA involves selecting one correct option from a set of four. We generate semantically preserving and inverting variations of these triplets to add more diversity for BQA.

Examples of generated data corresponding to each logic type and reasoning patterns are presented in App. C.

3.2.1 Sentence Generation

Here, the first step is to generate sentences with diverse *ontologies*. An ontology represents a collection of concepts (e.g. car, person, animals, etc.) along with their corresponding associated properties. To generate these sentences, we prompt the GPT-3.5 model with instructions tailored for each inference rule (more details in App. A).

An example of a prompt corresponding to the 'Modus Tollens' from PL is presented in App. A for better illustration. Note that our objective at this stage is not to generate logical sentences but rather to generate a diverse and coherent set of sentences that encompass various concepts. We also create a negation sentence corresponding to each generated sentence[2]. In this work, the scope of generating negations is simple (refer to App. C for examples), however, negations can be more complicated in the case of logic. These generated sentences will be combined with logical connectives in a later stage to form context and questions.

3.2.2 NL Conversion

Here, the NL conversion is accomplished using two steps. First, we leverage the formal expressions of reasoning patterns to create templates that establish the desired NL formulation for each logical connective (i.e., templatized context). Second, we prompt GPT-3.5 to transform the templatized context into a story/narrative-based context, enhancing its naturalness. For instance, implication: "$p \rightarrow q$" is expressed as "If p, then q", conjunction: "$p \wedge q$" as "p and q.", and disjunction: "$p \vee q$" as "At least one of the following is true: (1) p and (2) q. Note that we do not know which of (1) and (2) is true. It is possible that only (1) is true, or only (2) is true, or both are true." since understanding the logical implication of 'or' when integrated into logical formulations posed challenges to both humans and models. With these established formulations, we proceed to utilize the sentences generated in §3.2.1 to create the templatized context and questions corresponding to reasoning patterns.

Then, the templatized context is converted into a narrative-based context using a prompt-based converter, enhancing its naturalness. The prompt-based converter (essentially, prompting GPT-3.5) ensures that the context is no longer templatized yet follows the logical connection between sentences as mentioned in the logical rule (further details are presented in App. D). For instance, let's consider the "Modus Tollens" from PL $(((p \rightarrow$

[2]We use https://github.com/dmlls/negate to generate negated sentences

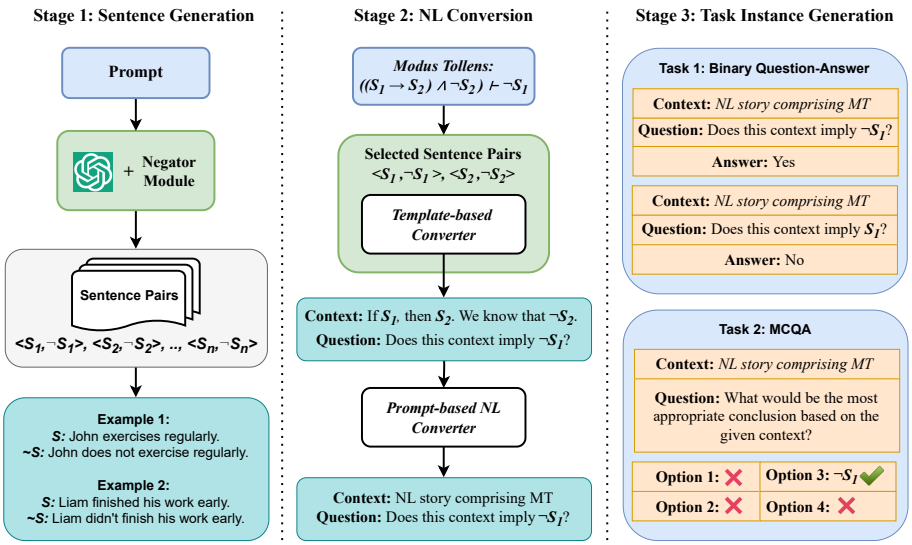

Figure 2: Schematic representation of three-stage procedure for data creation. NL: Natural Language

$q) \wedge \neg q) \vdash \neg p)$, and the "Bidirectional Dilemma" from FOL ($\forall x((p(x) \rightarrow q(x)) \wedge (r(x) \rightarrow s(x))) \wedge (p(a) \vee \neg s(a))) \vdash (q(a) \vee \neg r(a))$). For these rules, Table 4 presents examples of logical templatized and narrative-based context, question, and task instances for both, BQA and MCQA. App. D showcases further examples corresponding to each inference rule and patterns from *LogicBench*.

3.2.3 Task Instance Generation

After generating the context and questions in §3.2.2, we generate *(context, question, answer)* triplets for both tasks: (i) BQA, and (ii) MCQA. Here, narrative-based context is similar for both BQA and MCQA tasks, only the format of *(question, answer)* pairs are different.

BQA We generate semantically preserving and inverting variations of questions. Let's consider the example of "Modus Tollens" from Table 4, having question as: "Does this imply that Liam didn't finish his work early?" In this question, we observe one proposition: s_1, representing the statement "Liam didn't finish his work early," can be used to create another question that did not follow the logical rule MT. We can create two possible tuples: $< \neg s_1, yes >, < s_1, no >$. Each tuple has a question-answer combination using proposition s_1. Moreover, we do not generate variations for the context since it offers no substantial diversity in the dataset. For question variations, we replace the variation of proposition $< s_1 >$ in the original question with the corresponding tuples to add diversity to *LogicBench*. The process allows us to create more variations of the question for BQA tasks, as illustrated in Figure 2 (Step 3 - Task 1).

MCQA A prompt-based approach is used to create different incorrect options. For rule MT, as shown in Table 4, option 1 is a correct option that follows the rule MT logically ($< \neg s_1 >$), while the other three options are generated using prompting in a way that does not follow the rule being incorrect options. In addition to the options, the question is also replaced by a randomly selected question from the set of five questions. As seen in Figure 2 (Step 3 - Task 2), there is only one correct option out of four given options. More details related to incorrect option generation and instances examples for both tasks are in the App. C, and E.2.

3.3 Statistics and Qualitative Analysis

Statistics We introduce two versions of our proposed dataset: *LogicBench(Eval)*$_{BQA}$ and *LogicBench(Eval)*$_{MCQA}$. Statistics of both versions are presented in Table 5. For *LogicBench(Eval)*$_{BQA}$, out of 1520, 520 samples are for 'yes' and 1000 samples are for 'no' labels. For *LogicBench(Eval)*$_{MCQA}$, there are 20 unique samples present for each rule, thus in total 500 unique samples. Furthermore, we synthetically augmented *LogicBench(Eval)* for training purposes (i.e., *LogicBench(Aug)*) which consists of 150 unique data samples for each rule for BQA, resulting in a total of 12908 data samples including variations.

Data Validation Throughout the data generation phase of *LogicBench(Eval)*, the authors conduct

Generated Sentences in Stage 1	Context	Binary QA	MCQA
Inference rule: **MT** p: Liam finished his work early. ¬p: Liam did not finish his work early. q: he will order pizza for dinner. ¬q: he will not order pizza for dinner.	**Templatized Context:** If Liam finishes his work early, then he will order pizza for dinner. He won't order pizza for dinner. **NL Context:** Liam knows that if he finishes his work early for the day, he will order pizza for dinner. However, on this particular day, he decided against ordering pizza.	**Question 1:** Does this imply that Liam didn't finish his work early? (Yes) **Question 2:** Does this imply that Liam finishes his work early? (No)	**Question:** Based on the context, what conclusion would be deemed most suitable? **Option 1:** Liam didn't finish his work early. (Yes) **Option 2:** Sarah had already ordered Chinese takeout. (No) **Option 3:** Rebecca finished her work early. (No) **Option 4:** Liam decided to order sushi instead. (No)
Inference rule: **BD** p(x): someone drinks lots of water q(x): they will feel hydrated r(x): they eat too much sugar s(x): they will experience a sugar crash p(a): Jane drinks lots of water ¬p(a): Jane does not drink lots of water q(a): she will feel hydrated ¬q(a): she will not feel hydrated r(a): she eats too much sugar ¬r(a): she does not eat too much sugar s(a): she will experience a sugar crash ¬s(a): she will not experience a sugar crash	**Templatized Context:** If someone drinks lots of water, then they will feel hydrated. If they eat too much sugar, then they will experience a sugar crash. We know that at least one of the following is true (1) Jane drinks lots of water and (2) she won't experience a sugar crash. Note that we do not know which ones of (1) and (2) are true. It might be the case that only (1) is true, or only (2) is true or both are true. **NL Context:** If someone consumes a significant amount of water, they will experience a state of hydration. conversely, if excessive amounts of sugar are ingested by them, a sugar crash will ensue. it is known that at least one of the following statements is true: either the Jane consumes ample water or she will not experience a sugar crash. however, the actual veracity of either statement remains ambiguous, as it could be the case that only the first statement is true, only the second statement is true, or both statements are true.	**Question 1:** Can we say at least one of the following must always be true? (a) she will feel hydrated and (b) she doesn't eat too much sugar (Yes) **Question 2:** Can we say at least one of the following must always be true? (a) she won't feel hydrated and (b) she eats too much sugar (No) **Question 3:** Can we say at least one of the following must always be true? (a) she will feel hydrated and (b) she eats too much sugar (No) **Question 4:** Can we say at least one of the following must always be true? (a) she won't feel hydrated and (b) she doesn't eat too much sugar (No)	**Question:** Taking into account the context provided, what conclusion would be the most appropriate? **Option 1:** If Jane consumes ample water, she will experience a sugar crash (No) **Option 2:** John will feel hydrated or he won't experience a sugar crash (No) **Option 3:** Jane will feel hydrated or she doesn't eat too much sugar (Yes) **Option 4:** Jane won't feel hydrated or she will eat too much sugar (No)

Table 4: Illustrative examples of logical context and questions created using sentences that are generated in the first stage (§3.2.1) for both tasks, i.e., BQA and MCQA.

Dataset	# of Instances per Axiom	Total # of Instances	Total # of Instances (Including Variations)
LogicBench(Eval)$_{BQA}$	20	500	1520
LogicBench(Eval)$_{MCQA}$	20	500	500

Table 5: Statistics of *LogicBench(Eval)*

a review of the logical formations to ensure they follow the intended logical structure. We examine each narrative for any potential discrepancies, ensuring that they are logically sound and correctly represent the intended relationships between propositions. In addition to the logical formation, we also dedicated considerable effort to eliminating typos and validating the grammar. We also analyze the diversity in terms of different ontology and the logical nature of the *LogicBench(Eval)* (presented in App. C.1). We mitigate errors encountered during the validation step (presented in App. F).

4 Results and Analysis

4.1 Experimental Setup

Task Formulation For BQA, let us consider a set of data instances $\mathcal{I}_{r,L}$ corresponding to the inference rule r and logic type L. In this set, i^{th} instance is represented as $\mathcal{I}^i_{r,L} = \{(c_i, Q_i)\}$ where c_i represents narrative context and $Q_i = \{q_1, q_2, ..., q_n\}$ represents set of question and its variations corresponding to i^{th} instance. As discussed in §3, each context (c) represents logical rules (e.g., All cats have fur. Tom is a cat.) and question (q) represents the conclusion (e.g., Does Tom have fur?). To each context and question pair, i.e., $< c, q >$, we assign a label from the set $\mathcal{Y} = \{Yes, No\}$. We assign a label Yes if the conclusion logically entails the context, otherwise, assign a label No. To evaluate any LLMs on this setup, we provide $< p, c, q >$ as input to predict a label from $\mathcal{Y}$ where p is a natural language prompt. In the set $\mathcal{I}_{r,L}$ for MCQA, i^{th} instance is represented as $\mathcal{I}^i_{r,L} = \{(c_i, q_i, O_i)\}$ where c_i represents narrative context and q_i represents question and $O_i = \{o_1, o_2, o_3, o_4\}$ represents four option choices. To each context and question pair, i.e., $< c, q >$, we assign a label from the set $\mathcal{Y} = \{o_1, o_2, o_3, o_4\}$. We assign a label o_1 if the correct conclusion is presented in the first option, and likewise for other labels. To evaluate any LLMs on this setup, we provide $< p, c, q, o >$ as input to predict a label from $\mathcal{Y}$.

Experiments We evaluate a range of prompting models including GPT-4, ChatGPT (GPT-3.5-Turbo), Google Gemini-Pro, Llama-2-7B-Chat, and Mistral-7B-Instruct-v0.2. Each model is evaluated in a zero-shot setting where the chain-of-thought prompt is provided to the model without any in-context examples. This approach allows us to determine LLM's inherent ability to do logical reasoning (based on pre-training), as we can not expect that various logical inference rules/patterns will always be made part of prompts. However, we do evaluate these models in a few-shot setting, and present the results in App. G.

Metrics Here, we evaluate performance in terms of accuracy for both tasks, BQA and MCQA. For the BQA, we measure accuracy corresponding to each label, i.e., $A(Yes)$ and $A(No)$. We evaluate each model on three different chain-of-thought prompts and report average results across these prompts. All prompts used for experiments are described in App. H.

Type	Rules	Llama-2		Mistral		Gemini		ChatGPT		GPT-4	
		$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$
PL	HS	$100_{0.00}$	$47.35_{0.03}$	$98.63_{0.01}$	$64.24_{0.07}$	$99.31_{0.01}$	$59.60_{0.01}$	$100_{0.00}$	$66.82_{0.04}$	$100_{0.00}$	$84.75_{0.06}$
	DS	$44.81_{0.05}$	$56.82_{0.02}$	$50.77_{0.03}$	$77.66_{0.05}$	$68.21_{0.04}$	$91.65_{0.04}$	$51.26_{0.04}$	$80.71_{0.09}$	$79.96_{0.06}$	$100_{0.00}$
	CD	$79.94_{0.03}$	$25.47_{0.01}$	$83.37_{0.03}$	$72.30_{0.12}$	$85.56_{0.04}$	$28.15_{0.01}$	$90.75_{0.06}$	$38.63_{0.04}$	$92.85_{0.01}$	$66.12_{0.05}$
	DD	$82.22_{0.16}$	$25.22_{0.01}$	$71.26_{0.02}$	$14.16_{0.03}$	$75.48_{0.04}$	$25.22_{0.02}$	$71.28_{0.05}$	$23.47_{0.03}$	$84.51_{0.02}$	$42.40_{0.04}$
	BD	$48.89_{0.43}$	$25.13_{0.01}$	$79.68_{0.01}$	$47.36_{0.11}$	$86.81_{0.02}$	$29.51_{0.01}$	$84.91_{0.05}$	$33.47_{0.05}$	$87.86_{0.04}$	$59.06_{0.13}$
	MT	$70.54_{0.03}$	$71.46_{0.05}$	$47.96_{0.08}$	$44.96_{0.10}$	$75.99_{0.05}$	$81.47_{0.03}$	$55.63_{0.02}$	$66.11_{0.07}$	$55.28_{0.04}$	$59.05_{0.07}$
	MI	$78.57_{0.26}$	$25.34_{0.01}$	$75.36_{0.01}$	$25.51_{0.04}$	$74.69_{0.06}$	$24.84_{0.04}$	$81.60_{0.02}$	$31.79_{0.03}$	$91.84_{0.01}$	$39.72_{0.03}$
	CT	$70.00_{0.12}$	$24.98_{0.02}$	$87.38_{0.01}$	$71.99_{0.03}$	$88.31_{0.08}$	$35.11_{0.07}$	$89.88_{0.05}$	$43.33_{0.03}$	$98.59_{0.01}$	$60.71_{0.06}$
	Avg	$\mathbf{71.87_{0.13}}$	$\mathbf{37.72_{0.02}}$	$\mathbf{74.30_{0.03}}$	$\mathbf{52.27_{0.07}}$	$\mathbf{81.79_{0.04}}$	$\mathbf{46.94_{0.03}}$	$\mathbf{78.16_{0.04}}$	$\mathbf{48.04_{0.05}}$	$\mathbf{86.36_{0.03}}$	$\mathbf{63.98_{0.05}}$
FOL	EG	$100_{0.00}$	$71.67_{0.03}$	$100_{0.00}$	$100_{0.00}$	$85.68_{0.06}$	$97.97_{0.04}$	$96.74_{0.03}$	$96.75_{0.03}$	$98.41_{0.03}$	$100_{0.00}$
	UI	$80.24_{0.05}$	$56.04_{0.02}$	$85.24_{0.04}$	$85.24_{0.04}$	$90.69_{0.08}$	$90.29_{0.04}$	$85.31_{0.02}$	$96.19_{0.03}$	$90.16_{0.002}$	$91.58_{0.03}$
	MP	$97.44_{0.04}$	$73.01_{0.04}$	$96.28_{0.03}$	$95.24_{0.05}$	$98.25_{0.03}$	$96.97_{0.05}$	$100_{0.00}$	$91.03_{0.04}$	$94.74_{0.00}$	$100_{0.00}$
	HS	$100_{0.00}$	$36.83_{0.01}$	$98.02_{0.00}$	$65.00_{0.05}$	$97.70_{0.00}$	$52.46_{0.04}$	$95.14_{0.00}$	$46.74_{0.04}$	$89.03_{0.01}$	$61.21_{0.01}$
	DS	$58.33_{0.14}$	$52.08_{0.04}$	$61.69_{0.03}$	$92.59_{0.13}$	$75.96_{0.07}$	$85.82_{0.07}$	$72.46_{0.10}$	$94.71_{0.05}$	$63.34_{0.04}$	$100_{0.00}$
	CD	$100_{0.00}$	$26.36_{0.01}$	$82.21_{0.02}$	$82.02_{0.09}$	$94.71_{0.02}$	$33.62_{0.02}$	$87.87_{0.06}$	$39.37_{0.06}$	$92.65_{0.01}$	$74.72_{0.05}$
	DD	$54.37_{0.19}$	$23.36_{0.01}$	$73.78_{0.01}$	$6.67_{0.11}$	$82.70_{0.02}$	$28.96_{0.01}$	$72.70_{0.08}$	$26.42_{0.11}$	$85.54_{0.04}$	$54.55_{0.08}$
	BD	$92.31_{0.13}$	$25.78_{0.01}$	$75.76_{0.01}$	$44.44_{0.10}$	$82.82_{0.08}$	$28.93_{0.05}$	$85.50_{0.03}$	$38.82_{0.02}$	$84.22_{0.02}$	$66.67_{0.10}$
	MT	$77.11_{0.05}$	$88.41_{0.09}$	$69.01_{0.06}$	$91.71_{0.01}$	$71.18_{0.10}$	$89.58_{0.04}$	$62.46_{0.03}$	$92.80_{0.06}$	$64.69_{0.04}$	$100_{0.00}$
	Avg	$\mathbf{84.42_{0.07}}$	$\mathbf{50.39_{0.03}}$	$\mathbf{82.44_{0.02}}$	$\mathbf{73.66_{0.06}}$	$\mathbf{86.63_{0.05}}$	$\mathbf{67.17_{0.04}}$	$\mathbf{84.24_{0.04}}$	$\mathbf{69.20_{0.05}}$	$\mathbf{84.75_{0.02}}$	$\mathbf{83.19_{0.02}}$
NM	DRI	$29.06_{0.05}$	$40.98_{0.04}$	$61.51_{0.05}$	$72.22_{0.05}$	$65.45_{0.03}$	$78.50_{0.02}$	$55.77_{0.00}$	$88.89_{0.09}$	$81.29_{0.05}$	$100_{0.00}$
	DRS	$66.56_{0.03}$	$18.40_{0.01}$	$70.13_{0.01}$	$6.20_{0.01}$	$68.05_{0.01}$	$7.20_{0.06}$	$69.06_{0.01}$	$0.00_{0.00}$	$77.59_{0.02}$	$39.32_{0.06}$
	DRD	$49.21_{0.06}$	$50.47_{0.04}$	$77.64_{0.02}$	$97.78_{0.04}$	$68.10_{0.05}$	$97.43_{0.04}$	$55.11_{0.02}$	$100_{0.00}$	$86.96_{0.00}$	$100_{0.00}$
	DRO	$53.25_{0.01}$	$53.60_{0.02}$	$68.30_{0.04}$	$89.56_{0.11}$	$57.85_{0.04}$	$74.03_{0.08}$	$50.85_{0.01}$	$66.67_{0.58}$	$55.56_{0.00}$	$100_{0.00}$
	RE1	$94.07_{0.06}$	$27.55_{0.01}$	$78.79_{0.02}$	$44.39_{0.03}$	$84.04_{0.04}$	$41.98_{0.12}$	$75.82_{0.00}$	$31.74_{0.03}$	$85.44_{0.02}$	$100_{0.00}$
	RE2	$54.17_{0.47}$	$53.62_{0.05}$	$81.31_{0.04}$	$85.67_{0.03}$	$58.48_{0.03}$	$82.22_{0.17}$	$64.67_{0.11}$	$60.68_{0.02}$	$59.42_{0.00}$	$100_{0.00}$
	RE3	$39.05_{0.03}$	$38.05_{0.10}$	$78.61_{0.04}$	$83.59_{0.01}$	$64.67_{0.08}$	$79.68_{0.02}$	$59.05_{0.04}$	$83.57_{0.08}$	$83.12_{0.02}$	$89.08_{0.01}$
	RAP	$76.82_{0.05}$	$67.22_{0.03}$	$51.73_{0.02}$	$60.48_{0.11}$	$75.58_{0.04}$	$96.67_{0.06}$	$61.44_{0.02}$	$92.16_{0.03}$	$66.07_{0.05}$	$98.72_{0.02}$
	Avg	$\mathbf{57.78_{0.09}}$	$\mathbf{43.74_{0.04}}$	$\mathbf{71.00_{0.03}}$	$\mathbf{67.49_{0.05}}$	$\mathbf{67.78_{0.04}}$	$\mathbf{69.71_{0.07}}$	$\mathbf{61.47_{0.03}}$	$\mathbf{65.46_{0.10}}$	$\mathbf{74.43_{0.02}}$	$\mathbf{84.75_{0.02}}$

Table 6: Evaluation of LLMs in terms of label-wise accuracy on *LogicBench(Eval)*$_{BQA}$, where $A(Yes)$ and $A(No)$ denote the accuracy for the Yes and No labels, respectively. DRI: Default Reasoning with Irrelevant Information, DRS: Default Reasoning with Several Defaults, DRD: Default Reasoning with a Disabled Default, DRO: Default Reasoning in an Open Domain, RE1: Reasoning about Unknown Expectations I, RE2: Reasoning about Unknown Expectations II, RE3: Reasoning about Unknown Expectations III, RAP: Reasoning about Priorities

4.2 Main Results

Table 6 and Table 7 represent inference rule-wise performance corresponding to each LLMs for the BCQ and MCQA tasks, respectively. Specifically, Table 6 provides label-wise accuracy ($A(Yes)$ and $A(No)$) for the BQA task, and Table 7 provides overall accuracy for the MCQA task. Both tables provide valuable insights into the performance of different models on various logic types and lead to several interesting findings. From Table 6, we can observe that ChatGPT achieves 48.04%, and GPT-4 shows a performance of 63.98% $A(Yes)$ on average which indicates the challenge of classical logical reasoning (PL) even for larger LLMs such as ChatGPT and GPT-4. Furthermore, we can observe that models struggle more with inference rules of PL compared to FOL and NM. In addition, it is noticeable that each model performs relatively better on questions with a negative response (i.e., No) compared to questions with a positive response (i.e., Yes). This observation suggests that the models struggle to fully comprehend the logical relationship between the context and the conclusion (i.e., lower $A(Yes)$). However, they demonstrate a relatively stronger understanding when the relationship is contradictory in nature (i.e., higher $A(No)$). From Table 7, we can observe that larger models exhibit superior performance in selecting the correct choice to arrive at a logical conclusion. Interestingly, the performance of models decreases when the inference rules are longer or include negations. In contrast to Table 6, for MCQA, LLMs show superior performance for PL and FOL compared to NM. To further investigate these findings and provide a detailed analysis, we perform a thorough study of reasoning chains generated by LLMs and present our insights in the subsequent section.

4.3 Analysis and Discussion

Lower performance of LLMs on PL as compared to NM and FOL for BQA. In the development of AI, NM logic was partly developed to formalize natural language constructs, such as "normally birds fly", that were not formalizable in a straightforward manner using classical math-

Type	Rules	Llama-2	Mistral	Gemini	ChatGPT	GPT-4
PL	HS	$86.67_{0.08}$	$93.33_{0.06}$	$100_{0.00}$	$91.67_{0.03}$	$100_{0.00}$
	DS	$63.33_{0.12}$	$60.00_{0.09}$	$86.67_{0.10}$	$96.67_{0.06}$	$95.00_{0.00}$
	CD	$80.00_{0.05}$	$70.00_{0.13}$	$96.67_{0.03}$	$90.00_{0.00}$	$100_{0.00}$
	DD	$43.33_{0.03}$	$30.00_{0.10}$	$90.00_{0.05}$	$73.33_{0.06}$	$88.33_{0.03}$
	BD	$51.67_{0.03}$	$53.33_{0.03}$	$86.67_{0.03}$	$68.33_{0.06}$	$83.33_{0.03}$
	MT	$31.67_{0.03}$	$60.00_{0.05}$	$78.33_{0.08}$	$73.33_{0.08}$	$76.67_{0.03}$
	MI	$33.33_{0.13}$	$35.00_{0.05}$	$71.67_{0.06}$	$63.33_{0.08}$	$73.33_{0.14}$
	CT	$73.33_{0.10}$	$68.33_{0.08}$	$100_{0.00}$	$98.33_{0.03}$	$100_{0.00}$
	Avg	$\mathbf{57.92_{0.07}}$	$\mathbf{58.75_{0.07}}$	$\mathbf{88.75_{0.04}}$	$\mathbf{81.88_{0.05}}$	$\mathbf{89.58_{0.03}}$
FOL	EI	$80.00_{0.00}$	$85.00_{0.05}$	$95.00_{0.00}$	$93.33_{0.03}$	$100_{0.00}$
	UI	$63.33_{0.03}$	$75.00_{0.05}$	$98.33_{0.03}$	$91.67_{0.03}$	$98.33_{0.03}$
	MP	$85.00_{0.09}$	$98.33_{0.03}$	$100_{0.00}$	$100_{0.00}$	$100_{0.00}$
	HS	$61.67_{0.06}$	$70.00_{0.05}$	$81.67_{0.03}$	$73.33_{0.06}$	$76.67_{0.03}$
	DS	$43.33_{0.06}$	$36.67_{0.03}$	$70.00_{0.05}$	$78.33_{0.10}$	$95.00_{0.05}$
	CD	$75.00_{0.05}$	$61.67_{0.06}$	$93.33_{0.03}$	$80.00_{0.05}$	$91.67_{0.03}$
	DD	$36.67_{0.06}$	$46.67_{0.03}$	$85.00_{0.05}$	$71.67_{0.06}$	$93.33_{0.03}$
	BD	$35.00_{0.05}$	$43.33_{0.06}$	$78.33_{0.06}$	$66.67_{0.13}$	$91.67_{0.06}$
	MT	$41.67_{0.03}$	$66.67_{0.03}$	$81.67_{0.06}$	$86.67_{0.06}$	$86.67_{0.03}$
	Avg	$\mathbf{57.96_{0.05}}$	$\mathbf{64.81_{0.04}}$	$\mathbf{87.04_{0.03}}$	$\mathbf{82.41_{0.06}}$	$\mathbf{91.51_{0.04}}$
NM	DRI	$38.33_{0.03}$	$28.33_{0.06}$	$58.33_{0.08}$	$66.67_{0.06}$	$90.00_{0.00}$
	DRS	$41.67_{0.08}$	$16.67_{0.10}$	$45.00_{0.10}$	$41.67_{0.10}$	$55.00_{0.05}$
	DRD	$55.00_{0.00}$	$50.00_{0.05}$	$48.33_{0.03}$	$71.67_{0.10}$	$80.00_{0.05}$
	DRO	$21.67_{0.03}$	$21.67_{0.03}$	$53.33_{0.03}$	$38.33_{0.08}$	$45.00_{0.00}$
	RE1	$51.67_{0.03}$	$31.67_{0.08}$	$70.00_{0.00}$	$65.00_{0.05}$	$95.00_{0.05}$
	RE2	$65.00_{0.05}$	$75.00_{0.00}$	$68.33_{0.08}$	$61.67_{0.03}$	$66.67_{0.06}$
	RE3	$31.67_{0.03}$	$33.33_{0.03}$	$61.67_{0.08}$	$70.00_{0.05}$	$68.33_{0.03}$
	RAP	$46.67_{0.08}$	$35.00_{0.09}$	$33.33_{0.03}$	$55.00_{0.05}$	$51.67_{0.03}$
	Avg	$\mathbf{43.96_{0.04}}$	$\mathbf{36.46_{0.05}}$	$\mathbf{54.79_{0.05}}$	$\mathbf{58.75_{0.07}}$	$\mathbf{68.96_{0.03}}$

Table 7: Evaluation of LLMs in terms of accuracy on *LogicBench(Eval)*$_{MCQA}$.

ematical logics. Thus, while it was difficult for researchers to come up with non-monotonic logics and formalize non-monotonic reasoning, the fact that they were usually motivated by natural language examples, suggests that many of the non-monotonic reasoning aspects are present in the NL text in the wild that is used in the pre-training of the ultra-large LLMs such as GPT4. While from human experience and complexity theory, FOL is harder than PL in general; in the LLM context, the crucial factor becomes what kind of logical sentences LLMs are pre-trained on. It seems that LLMs are pre-trained more on simple FOL sentences than on simple PL sentences (see Appendix I for further discussion). On the other hand, some PL features are perhaps less prevalent in human writing (on which LLMs are pre-trained) - such as Modes Tollens. Table 6 shows that GPT-4 achieves $\sim 85\%$ accuracy ($A(Yes)$) for simple inference rules such as HS(PL). However, GPT-4 performance dropped to $\sim 59\%$ $A(Yes)$ for PL(MT).

Negations are hard to understand when embedded with logical rules. Regarding PL and FOL, it is apparent that the models struggle more with the DD and MT inference rules. A closer look at Table 2 reveals that all of these rules include examples where the models need to draw conclusions based on negated premises. This indicates that the models encounter difficulties when negated premises are introduced. We also analyze the effect of negations on the reasoning chain (see App. I).

Longer inference rules are still challenging. Table 6 indicates that the models face challenges when handling longer rules, such as BD, CD, and DD, both in PL and FOL. Hence, it can be concluded that these models struggle with longer logical dependencies in the premise, particularly when a higher number of propositions are present. In the case of NM reasoning, the models exhibit lower performance in DRS, indicating that a higher number of premises often leads to more frequent mistakes.

Lower performance of NM compared to PL for MCQA. We examine the reasoning chains generated for instances sharing the same contexts in both BQA and MCQA tasks, focusing on PL and NM. For PL, we note that LLMs tend to rely on parametric knowledge in BQA task due to the absence of choices, yet they demonstrate good performance in MCQA tasks when choices are available. Conversely, NM exhibits contrasting behavior. For a more in-depth analysis, please refer to App. I.

Effect on other logic datasets We trained the T5-large model on the *LogicBench(Aug)* resulting in a model named LogicT5. Furthermore, we performed fine-tuning on four other logical reasoning datasets: LogiQA, Reclor, LogicNLI, and FOLIO. Further discussion is presented in App. I.

5 Conclusions

In this work, we evaluated the logical reasoning ability of LLMs on 25 distinct inference rules and reasoning patterns covering PL, FOL, and NM logics. To this end, we introduced *LogicBench*, a natural language question-answering dataset focusing on evaluating a single inference rule. We devised two tasks using *LogicBench*: (i) BQA, and (ii) MCQA. We evaluated a range of LLMs including GPT-4, ChatGPT, Gemini-Pro, Llama-2, and Mistral on both tasks. Experimental results showed that LLMs do not perform well on *LogicBench*, even though they require the application of only a single inference rule. Furthermore, we also augmented *LogicBench* to *LogicBench(Aug)*, which can be utilized for training purposes. Using *LogicBench(Aug)*, we demonstrated that LLMs trained using it showcase an improved understanding of logical reasoning, resulting in a better performance on existing logic datasets.

Limitations

While *LogicBench* encompasses 25 distinct inference rules spanning three logic types (significantly more than any previous study) to comprehensively evaluate the logical reasoning capabilities of LLMs, it can be further extended by incorporating additional inference rules and logic types. However, with respect to first-order logic and logics with quantified variables, there can be an infinite number of such rules. In this study, we focused solely on evaluating model performance using a single inference rule; however, an interesting future direction can be enhancing the depth of reasoning complexity by incorporating combinations of inference rules to derive conclusions. 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.

References

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2020. Fracas: Temporal analysis.

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.

Xiang Chen, Xin Xie, Ningyu Zhang, Jiahuan Yan, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Adaprompt: Adaptive prompt-based finetuning for relation extraction. *arXiv e-prints*, pages arXiv–2104.

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.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using bart. *arXiv preprint arXiv:2106.01760*.

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

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Himanshu Gupta, Neeraj Varshney, Swaroop Mishra, Kuntal Kumar Pal, Saurabh Arjun Sawant, Kevin Scaria, Siddharth Goyal, and Chitta Baral. 2022. " john is 50 years old, can his son be 65?" evaluating nlp models' understanding of feasibility. *arXiv preprint arXiv:2210.07471*.

Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2021. Towards general purpose vision systems. *arXiv preprint arXiv:2104.00743*.

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

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

Peter Hase and Mohit Bansal. 2021. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*.

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*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on*

Computational Natural Language Learning, pages 41–55, Online. Association for Computational Linguistics.

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

Kirby Kuznia, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Less is more: Summary of long instructions is better for program synthesis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4532–4552, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press, Rome, Italy.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021a. 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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Man Luo, Sharad Saxena, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Biotabqa: Instruction learning for biomedical table question answering. *arXiv preprint arXiv:2207.02419*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. The SICK (Sentences Involving Compositional Knowledge) dataset for relatedness and entailment.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021a. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021b. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

Swaroop Mishra and Elnaz Nouri. 2022. Help me think: A simple prompting strategy for non-experts to create customized content with models. *arXiv preprint arXiv:2208.08232*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. In-BoXBART: Get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Is a question decomposition unit all we need? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4553–4569, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *ICLR*.

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*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7063–7071.

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.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. QuaRTz: An open-domain dataset of qualitative relationship questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Cad: the contextual abuse dataset.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022a. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint arXiv:2203.03903*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. 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.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E Peters. 2020. Learning from task descriptions. *arXiv preprint arXiv:2011.08115*.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning?

Qinyuan Ye and Xiang Ren. 2021. Zero-shot learning by generating task-specific adapters. *arXiv e-prints*, pages arXiv–2101.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

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*.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Example Prompt for Sentence Generation

Below is the general prompt structure prompted to generate data. The prompt schema, as depicted in Figure 3, comprise three crucial components:

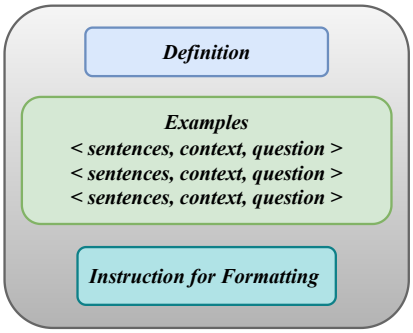

Figure 3: Schematic representation of prompt.

Definition provides a detailed explanation of the task and offers a natural language representation of the reasoning pattern for which we are generating sentences.

Examples provide sample sentences that need to be generated. We also illustrate how these sentences will be utilized in later stages, emphasizing the importance of coherence and the inclusion of relevant ontological concepts.

Format We provide specific formatting instructions to guide the generation of sentences.

Figure 4 illustrates an example prompt for the inference rule, namely, 'modus tollens' from propositional logic (PL). Modus tollens is formally represented as $((p \rightarrow q) \wedge \neg q) \vdash \neg p$, which can be understood in natural language as "If p implies q, and we know $\neg q$, then we can conclude $\neg p$." In this prompt, the definition provides a comprehensive description of the inference rule in natural language. To encourage the generation of more relevant and coherent sentences, the prompt includes an examples section that demonstrates how the generated sentences will be utilized in a later stage. This serves, as an illustration, to guide GPT-3 in producing suitable outputs. In Figure 4, we present three examples involving sentences p and q, along with their respective contexts and questions. The prompt also includes instructions on how the generated sentences should be formatted.

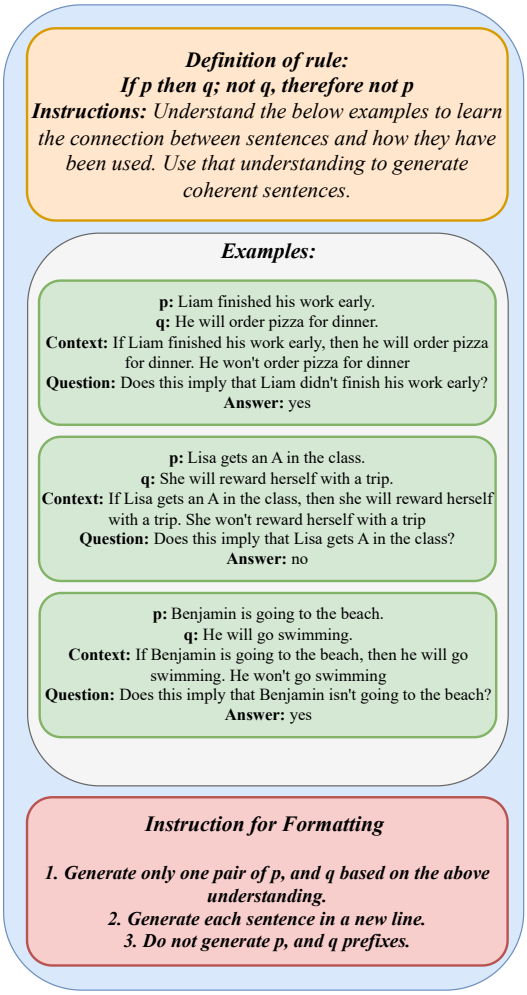

Figure 4: Example prompt for *Modus Tollens* inference rule from PL.

B Extended Related Work

As LLMs such as GPT-4, and Bard continue to evolve rapidly, it becomes increasingly crucial to evaluate their diverse language capabilities, as well as those of forthcoming LLMs. Recently, many datasets have been created that evaluate different language understanding skills such as pronoun resolution (Sakaguchi et al., 2021; Levesque et al., 2012), commonsense reasoning (Talmor et al., 2019), numerical reasoning (Dua et al., 2019; Patel et al., 2021; Mishra et al., 2022), qualitative reasoning (Tafjord et al., 2019b,a), temporal reasoning (Zhou et al., 2019), and feasibility reasoning (Gupta et al., 2022). Now, we present the advancements in prompt and instruction tuning using LLMs.

Prompt Learning The introduction of LLMs has significantly shifted the research trend in NLP to prompt-based learning methodologies (Liu et al., 2021b). Many studies have been conducted to investigate the efficacy of prompt-based learning in various applications including Text classification (Yin et al., 2019), Natural Language Inference

(NLI) (Schick and Schütze, 2020), and Question Answering (QA) (Jiang et al., 2020), Information Extraction (IE) (Chen et al., 2021; Cui et al., 2021), to name a few. In a recent development, the T0 model employs prompts to achieve zero-shot generalization across various NLP tasks (Sanh et al., 2021). Scao et al. 2021 suggested that the use of prompts could be as valuable as hundreds of data points on average (Le Scao and Rush, 2021).

Instruction Learning Efrat et al., 2020 (Efrat and Levy, 2020) was focused on whether existing LLMs understand instructions. The same work in the field of instruction by (Hase and Bansal, 2021; Ye and Ren, 2021; Gupta et al., 2021; Zhong et al., 2021) has been proposed to show that models follow natural language instructions. In addition, Weller et al., 2020 (Weller et al., 2020) developed a framework focusing on NLP systems that solve challenging new tasks based on their description. Mishra et al., 2021 (Mishra et al., 2021b) have proposed natural language instructions for cross-task generalization of LLMs. Similarly, PromptSource (Sanh et al., 2021) and FLAN (Wei et al., 2021) were built for leveraging instructions and achieving zero-shot generalization on unseen tasks. Moreover, Parmar et al., 2022 (Parmar et al., 2022) shows the effectiveness of instructions in multi-task settings for the biomedical domain. Furthermore, Mishra et al., 2021 (Mishra et al., 2021a) discuss the impact of task instruction reframing. Min et al., 2021 (Min et al., 2021) introduce a framework to better understand in-context learning. Ouyang et al., 2022 (Ouyang et al., 2022) propose the InstructGPT model that is fine-tuned with human feedback to follow instructions. Wang et al., 2022 (Wang et al., 2022a) has developed an instruction-based multi-task framework for few-shot Named Entity Recognition (NER) tasks. In addition, many approaches have been proposed to improve model performance using instructions (Wu et al., 2022; Lin et al., 2021; Wang et al., 2022b; Luo et al., 2022; Kuznia et al., 2022; Patel et al., 2022; Mishra and Nouri, 2022).

Logic and NLI Datasets FraCas (Bernardy and Chatzikyriakidis, 2020) offers a unique approach to temporal semantics by converting syntax trees into logical formulas tailored for inference, emphasizing temporal elements such as references, adverbs, aspectual classes, and progressives. The Monotonicity Entailment Dataset (MED) (Yanaka et al., 2019) dives deep into monotonicity reasoning within NLI, probing the synergy between lexical and syntactic structures and spotlighting inherent challenges in both upward and downward monotonic reasoning trajectories. The SICK (Marelli et al., 2014) dataset, with its foundation in 10,000 English sentence pairs, is designed to rigorously evaluate semantic relatedness and entailment, leveraging crowdsourced annotations for precision. HANS, or Heuristic Analysis for NLI Systems (McCoy et al., 2019), stands out by rigorously scrutinizing the dependability of NLI models, putting the spotlight on potential pitfalls tied to syntactic heuristics such as lexical overlap. Lastly, CAD (Vidgen et al., 2021) introduces a meticulously crafted dataset from Reddit entries, targeting the detection of online abuse. This dataset boasts six distinct primary categories, context-aware annotations, provided rationales, and a rigorous group-adjudication methodology ensuring high-quality annotations.

C Examples of Data Instances

This section provides examples of *(context, question, answer)* tuples corresponding to each inference rule and reasoning pattern. Additionally, it highlights the diverse range of question variations within the dataset associated with each inference rule and reasoning pattern.

C.1 Word Cloud

Figure 5 provides a word cloud derived from the $LogicBench(Eval)$. This word cloud highlights the logical nature and diversity of our evaluation dataset. Words such as 'if', 'normally', 'usually', and 'then' are prominently featured, suggesting their frequent use in the dataset, and suggesting the logical nature of the dataset. Moreover, we can also observe several words consisting of different ontologies such as 'cat', 'car', 'garden', and many more, suggesting diversity in the dataset.

C.2 Propositional Logic (PL)

Here, we discuss examples of each inference rule present in the PL of the *LogicBench* as shown in Table 8. Table 8 has context related to the inference rule and different variations of the question according to the rule. For instance, the first row of Table 8 shows the example for inference rule, *Hypothetical Syllogism (HS)*, formally expressed as $((p \rightarrow q)) \wedge (q \rightarrow r)) \vdash (p \rightarrow r)$. The context represents the premise, i.e., $((p \rightarrow q)) \wedge (q \rightarrow r))$,

Figure 5: Word cloud of context present in the LB(eval)

and the first question (Q1) represents the conclusion, i.e., $p \rightarrow r$. Hence, Q1 is labeled as "Yes" since it supports the conclusion given the logical context. Furthermore, Q2 to Q4 represent different variations of the question by utilizing the variables $(p, \neg p, r, \neg r)$. For the HS, given the provided context, Q2 to Q4 contain the variations $\neg p \rightarrow r$, $p \rightarrow \neg r$, and $\neg p \rightarrow \neg r$, respectively, and are labeled as "No" since they do not support the conclusion.

C.3 First-Order Logic (FOL)

Here, we discuss examples of each inference rule and two axioms (i.e., Existential Instantiation and Universal Instantiation) present in the FOL from the *LogicBench* as shown in Table 9. *Existential Generalization (EG)*, formally expressed as $P(a) \Rightarrow \exists x P(x)$ indicates that there is an element a in the domain for which $P(a)$ is true, then we know that $\exists x P(x)$ is true. *Universal Instantiation* formally expressed as $\forall x\, A \Rightarrow A\{x \mapsto a\}$ indicates that a statement holds true for all instances (x) within a specific category A, hence it is also true for specific instance a.

Table 9 represents context related to the inference rule and variations of the question. The process of generating data instances for FOL follows a similar approach to that of PL. For example, the first row of Table 9 shows the example for axiom, *Existential Instantiation (EI)*, formally expressed as $\exists x P(x) \Rightarrow P(a)$. The context represents the initial premise $\exists x P(x)$ and the first question (Q1) represents the conclusion, i.e., $P(a)$. Hence, Q1 is labeled as "Yes" since it supports the conclusion given the logical context. Furthermore, we generate the only variant of the question based on $\neg P(a)$ and labeled it as No since it does not support the conclusion.

C.4 Non-Monotonic (NM) Reasoning

Here, we discuss examples of each reasoning pattern present in the NM reasoning from the *LogicBench* as shown in Table 10. Table 10 has context related to the reasoning pattern and different variants of the question. For example, the first row of Table 10 shows the example for *Default Reasoning with Irrelevant Information (DRI)*. For this reasoning, based on the given context, there are also two possible variations of the question where one with a correct conclusion labeled as Yes and another with an incorrect conclusion labeled as No.

D Examples of NL Conversion

This section illustrates the way natural language logical context and questions are created using the generated sentence in Stage 1 in addition to prompt-based templatized-context to narrative conversion. Table 12 shows examples of how context and question are generated from sentences corresponding to each inference rule for PL and FOL. Similarly, Table 13 shows examples of NM reasoning. From Table 12, we can see an example of sentence pairs (p, q) and their corresponding negation pairs $(\neg p, \neg q)$ for the 'modus tollens' inference rule from PL. These pairs are utilized to generate templatized logical context which is converted to narrative using a specific prompt[1D] designed for PL. In addition to the narrative, the question and its variations are also created using these sentence pairs. Similarly, in the second row, we have four generic rules with variable x $(p(x), q(x), r(x), s(x))$ and their specific cases (i.e., $x = a$), along with their respective negative sentence pairs $[(p(a), \neg p(a))$, $(q(a), \neg q(a))$, $(r(a), \neg r(a))$, $(s(a), \neg s(a))]$. In the same manner, as PL, these examples demonstrate the generation of narrative and questions for the FOL inference rule called 'Bidirectional Dilemma (BD)', as shown in Table 12 using specific prompt[2D] designed for FOL. From Table 13, the first row presents an example of narrative and questions generated from a sentence pair for the 'Default Reasoning with Irrelevant Information (DRI)' from NM reasoning. In this specific instance, the generated sentences are (p, q, r, s, t), and the negation is only required for the sentence t. Therefore, there is a single negation pair $(t, \neg t)$, which is used to generate questions specific to the 'DRI'. Same as PL and FOL, templatized context is also converted to narrative using a prompt[3D] created for NM.

Rule	Context	Question
HS	If Jim cleans his room, he will receive a reward and if he receives a reward, he will use it to buy a new toy. So, Jim decided to tidy up his room, hoping to earn a reward. He diligently gathered his clothes, organized his toys, and dusted every surface. After a few hours of hard work, Jim's room was spotless.	**Q1:** If Jim cleaned his room, does this imply that he will buy a new toy? (Yes) **Q2:** If Jim didn't clean his room, does this entail that he won't buy a new toy? (No) **Q3:** If Jim cleaned his room, does this imply that he won't buy a new toy? (No) **Q4:** If Jim didn't clean his room, does this imply that he will buy a new toy? (No)
DS	Either Chloe is studying for her exams, or Mila is going on vacation, or both scenarios are unfolding. It is unclear which of the two options is true at this point. However, one thing is certain - Chloe is not studying for her exams.	**Q1:** Does this entail that Mila is going on vacation? (Yes) **Q2:** Does this mean that Mila isn't going on vacation? (No)
CD	If I decide to go for a walk, I will be able to breathe in some fresh air and revitalize myself. On the other hand, if I choose to stay home, I will have the opportunity to enjoy a movie. One thing is certain, either I go for a walk or I stay home. It remains uncertain which of the two options I will ultimately choose. It is entirely possible that I might opt for the walk, or perhaps I will find myself drawn to staying home, or even both possibilities might come to fruition.	Can we say at least one of the following must always be true? **Q1:** (a) I will get some fresh air or (b) I will watch a movie (Yes) **Q2:** (a) I won't get some fresh air and (b) I will watch a movie (No) **Q3:** (a) I will get some fresh air and (b) I won't watch a movie (No) **Q4:** (a) I won't get some fresh air and (b) I won't watch a movie (No)
DD	If I decide to order takeout, it means I will save time. On the other hand, if I choose to cook a meal, it means I will save money. The interesting thing is that I am in a situation where I won't be able to save time or money. It is uncertain whether I won't save time or I won't save money, or it could even be both scenarios. The only thing that is clear is that at least one of these possibilities is true.	Can we say at least one of the following must always be true? **Q1:** (a) I don't order takeout or (b) I don't cook a meal (Yes) **Q2:** (a) I order takeout and (b) I cook a meal (No) **Q3:** (a) I don't order takeout and (b) I cook a meal (No) **Q4:** (a) I order takeout and (b) I don't cook a meal (No)
BD	If it is sunny outside, then I will go for a walk. However, if it rains, I will stay inside. Currently, it is uncertain whether it is raining or not, but I do know that at least one of the following is true: either it is raining or I will not go for a walk. It is possible that only one of these statements is true, or perhaps both are true.	Can we say at least one of the following must always be true? **Q1:** (a) we will stay inside or (b) it is not sunny (Yes) **Q2:** (a) we will not stay inside and (b) it is sunny (No) **Q3:** (a) we will stay inside and (b) it is sunny (No) **Q4:** (a) we will not stay inside and (b)it is not sunny (No)
MT	If Mason decides to leave his job, he will not receive any salary. However, against all odds, Mason still receives his salary. He finds himself receiving his regular paycheck.	**Q1:** Does this infer that Mason didn't leave his job? (Yes) **Q2:** Does this infer that Mason left his job? (No)
MI	Rohan woke up in the morning and realized that he had forgotten his lunch. Knowing that if he forgets his lunch, he will not eat at school, he felt disappointed.	Based on context, can we say, at least one of the following must always be true? **Q1:** (a) Rohan didn't forget his lunch and (b) he will not eat at school (Yes) **Q2:** (a) Rohan forgot his lunch and (b) he will eat at school (No) **Q3:** (a) Rohan forgot his lunch and (b) he will not eat at school (No) **Q4:** (a) Rohan didn't forget his lunch and (b) he will eat at school (No)
CT	At least one of two things is true about Tom - he is either an avid reader or he devours books of all genres. We are unsure which one of these statements is true or if both are true. It could be that only the first statement is true, or only the second statement is true, or even that both are true.	Can we say at least one of the following must always be true? **Q1:** (a) he devours books of all genres or (b) Tom is an avid reader (Yes) **Q2:** (a) he doesn't devour books of all genres and (b) Tom is an avid reader (No) **Q3:** (a) he devours books of all genres and (b) Tom isn't an avid reader (No) **Q4:** (a) he doesn't devour books of all genres and (b) Tom isn't an avid reader (No)

Table 8: Examples of context and question-answer pairs for each rule of *Proportional logic* from the LogicBench; HS: Hypothetical Syllogism, DS: Disjunctive Syllogism, CD: Constructive Dilemma, DD: Destructive Dilemma, BD: Bidirectional Dilemma, MT: Modus Tollens, MI: Material Implication, CT: Commutation

Rule	Context	Question
UI	All students are required to take an examination in order to fulfill the requirements for their degree. Reema, being a student, is also expected to fulfill the requirements.	**Q1:** Does Reema need to take an exam to complete her degree? (Yes) **Q2:** Does Reema need not to take an exam to complete her degree? (No)
EG	The marathon race was won by James, who emerged as the champion.	**Q1:** Does this imply that someone won the marathon race? (Yes) **Q2:** Does this mean that no one won the marathon race? (No)
MP	If someone is extremely tired, then they will seek some rest and relaxation. Today, Jack finds himself utterly exhausted.	**Q1:** Does this entail that he will take rest? (Yes) **Q2:** Does this entail that he won't take rest? (No)
HS	If all the necessary supplies have been purchased by someone, then they can initiate the project. Once the project is started by someone, they will complete it within the expected time-frame.	**Q1:** If Lily bought all the necessary supplies, does this mean that she will finish it on time? (Yes) **Q2:** If Lily didn't buy all the necessary supplies, does this imply that she won't finish it on time? (No) **Q3:** If Lily bought all the necessary supplies, does this entail that she won't finish it on time? (No) **Q4:** If Lily didn't buy all the necessary supplies, does this imply that she will finish it on time? (No)
DS	It is known that one of the following options is true: someone goes to a museum or someone visits a park. The specific scenario could involve only the option to go to a museum being true, or only the option to visit a park being true, or both options being true. However, it is stated that Jill is unable to go to a museum.	**Q1:** Does this imply that she can visit a park? (Yes) **Q2:** Does this entail that she can't visit a park? (No)
CD	If someone is painting a picture, then they will frame it. similarly, the natural course of action for a writer would be to publish their completed story. in this scenario, it is certain that at least one of the following statements holds true: (1) john is currently engrossed in painting a picture, or (2) john is currently immersed in the act of writing a story. it should be emphasized that we are unaware of which statement specifically applies, as there is a possibility that either (1) alone is true, or (2) alone is true, or even that both (1) and (2) are simultaneously true.	Can we say at least one of the following must always be true? **Q1:** (a) he will frame it and (b) he will publish it. (Yes) **Q2:** (a) he won't frame it and (b) he will publish it. (No) **Q3:** (a) he will frame it and (b) he won't publish it. (No) **Q4:** (a) he won't frame it and (b) he won't publish it. (No)
DD	If someone is taking care of their health, then they will be fit. However, indulging in unhealthy habits can make individuals susceptible to various diseases. The truth is, we can be certain about at least one of the following possibilities: either Jenny won't be fit and healthy, or she won't be prone to diseases. It is important to note that we are unaware of which statement is accurate. It could be the case that only the first statement is true, only the second statement is true, or both statements hold validity.	Can we say at least one of the following must always be true? **Q1:** (a) Jenny doesn't take care of her health and (b) she doesn't indulge in unhealthy habits (Yes) **Q2:** (a) Jenny takes care of her health and (b) she indulges in unhealthy habits (No) **Q3:** (a) Jenny doesn't take care of her health and (b) she indulges in unhealthy habits (No) **Q4:** (a) Jenny takes care of her health and (b) she doesn't indulge in unhealthy habits (No)
BD	If an individual consumes a significant amount of water, they will experience a state of hydration. Conversely, if excessive amounts of sugar are ingested, a sugar crash will ensue. It is known that at least one of the following statements is true: either the Jane consumes ample water or she will not experience a sugar crash. However, the actual veracity of either statement remains ambiguous, as it could be the case that only the first statement is true, only the second statement is true, or both statements are true.	Can we say at least one of the following must always be true? **Q1:** (a) she will feel hydrated and (b) she doesn't eat too much sugar (Yes) **Q2:** (a) she won't feel hydrated and (b) she eats too much sugar (No) **Q3:** (a) she will feel hydrated and (b) she eats too much sugar (No) **Q4:** (a) she won't feel hydrated and (b) she doesn't eat too much sugar (No)
MT	If someone decides to go to the park, it is required that they wear a mask. However, in this particular situation, John does not wear a face covering.	**Q1:** Does this imply that John doesn't visit the park? (Yes) **Q2:** Does this entail that John visits the park? (No)

Table 9: Examples of context and question-answer pairs for each rule of *First order logic* from the LogicBench.

Rule	Context	Question
DRI	Once upon a time, in a land filled with animals, there were two popular mammalian creatures, cats and dogs. Mammals typically possessed a coat of fur, which kept them warm and protected. However, cats were an exception to this rule, as their bodies lacked fur. Nonetheless, both cats and dogs were beloved by many for their unique traits. Dogs, known for their loyalty, were particularly cherished by humans.	**Q1:** Does this imply that dogs have fur? (Yes) **Q2:** Does this entail that dogs don't have fur? (No)
DRS	John and Mary were expecting their first child, filled with the anticipation and excitement that all parents feel. Parents are usually loving and supportive. Parents are normally responsible. However, something seemed amiss in their relationship. Mary, usually affectionate and caring, seemed distant and uninvolved. On the other hand, John, known for his responsible nature, started neglecting his duties and became unreliable.	**Q1:** Does this imply that Mary is responsible and John is loving and supportive? (Yes) **Q2:** Does this entail that Mary isn't responsible and John is loving and supportive? (No) **Q3:** Does this imply that Mary is responsible and John isn't loving and supportive? (No) **Q4:** Does this entail that Mary isn't responsible and John isn't loving and supportive? (No)
DRD	Jenny and Anna are known for their tall stature, which is often associated with playing basketball. However, Anna might be an exception to this norm.	**Q1:** Does this entail that Jenny plays basketball? (Yes) **Q2:** Does this mean that Jenny doesn't play basketball? (No)
DRO	In the bird kingdom, there are many different species that possess unique characteristics. One such species is the hummingbird, known for its ability to hover in mid-air and its vibrant colors. While most birds engage in the annual migration south for the winter, the hummingbird chooses to stay put and brave the cold weather. This decision sets the hummingbird apart from its fellow avian companions, as it relies on its resilience and resourcefulness to survive the harsh conditions.	**Q1:** Does this mean that all other birds than hummingbirds migrate south for the winter? (Yes) **Q2:** Does this mean that all other birds than hummingbirds don't migrate south for the winter? (No)
RE1	In a world where animals are often regarded as intelligent creatures, there is a captivating tale that revolves around cats, dogs, and horses. It is commonly believed that most animals possess a level of intellect. However, there is an intriguing twist to this belief as it is known that either cats or dogs are not considered particularly intelligent. As the story unfolds, we delve into the lives of these remarkable creatures, their interactions, and the unique qualities that each of them possesses.	**Q1:** Does this entail that horses are considered to be intelligent creatures and exactly one of the cats or dogs is not considered intelligent? (Yes) **Q2:** Does this mean that horses aren't considered to be intelligent creatures and exactly one of cats or dogs is not considered intelligent? (No) **Q3:** Does this mean that horses are considered to be intelligent creatures and exactly one of cats or dogs is considered intelligent? (No) **Q4:** Does this implies that horses aren't considered to be intelligent creatures and exactly one of cats or dogs is considered intelligent? (No)
RE2	In the realm of cat communication, meowing serves as a fundamental aspect of their vocal repertoire. However, intriguingly enough, there exists a distinct species of cat that deviates from this conventional norm. This peculiar feline defies the expectations associated with its kind by refraining from emitting any meows whatsoever.	**Q1:** Does this entail that exactly one species of cat doesn't meow? (Yes) **Q2:** Does this imply that exactly one species of cat meows? (No)
RE3	In a world where cars were known for having four wheels, it was considered a common fact that wheels typically came equipped with spokes. However, amidst this widespread understanding, there was an exception. At least one wheel defied this norm and stood out from the rest by not having any spokes at all.	**Q1:** Does this imply that cars have four wheels with spokes? (Yes) **Q2:** Does this mean that cars don't have four wheels with spokes? (No)
RAP	In the midst of a heated argument, John adamantly claims that sally was present at the store. However, Jane strongly opposes John's assertion, insisting that Sally was indeed absent from the store.	**Q1:** If John's evidence is more reliable than Jane's, does this mean that Sally was in the store? (Yes) **Q2:** If John's evidence is more reliable than Jane's, does this mean that Sally wasn't in the store? (No) **Q3:** If John's evidence is less reliable than Jane's, does this entail that Sally was in the store? (No) **Q4:** If John's evidence is less reliable than Jane's, does this imply that Sally wasn't in the store? (Yes)

Table 10: Examples of context and question-answer pairs for each rule of *Non-monotonic logic* from the LogicBench.

Each prompt is shown below for PL, FOL, and NM templatized context to narrative conversion.

1. Prompt for templatized to narrative conversion for PL(MT):

> **Rule: Condition:** If p then q; **Situation:** not q
> **Context:** Templatized context
> Improve the context in human-like language and make story with rephrased sentences.
> **Instructions for generating a good story:**
> 1. While generating a story, Use the rephrased sentences from the context of the story.
> 2. Make sure to include sentences corresponding to the condition and situation from the rule in the story.
> 3. Do not add any other extra information.
> 4. To generate a story, DO NOT change the name of the main character from the context, if there's any. If the main character within the context is 'I', please generate a story with 'I' as the main character.
> 5. Generate only one paragraph with rephrased sentences.

2. Prompt for templatized to narrative conversion for FOL:

> **Context:** Templatized context
> Improve the context in human-like language and make a rephrase sentences around it.
> **Instructions for generating a good story:**
> 1. Make sure to include only rephrased sentences in the story.
> 2. Do not add any other extra information and do not reach any conclusion based on context.
> 3. To generate a story, use only generic sentences without mentioning any main character.
> 4. Extract a specific case sentence from the context, rephrase it, and append it at the end of the context only.
> 5. Generate only one paragraph with rephrased sentences.
> 6. DO NOT add any prefixes.

3. Prompt for templatized to narrative conversion for NM:

> **Context:** Templatized context
> Improve the context in human-like language and rephrase sentences around it.
> **Instructions for generating a good story:**
> 1. Make sure to include only rephrased sentences in the story.
> 2. Do not add any other extra information, but add some story-like context.
> 3. To generate a story, DO NOT change the name of the main character from the context, if there's any. Use the same names/words for main character. If the main character within the context is 'I', please generate a story with 'I' as the main character.
> 4. Generate only one paragraph with rephrased sentences.

To ensure the quality of the narrative in the *LogicBench(Eval)* for task 1 and task 2, we have created category-specific prompts to convert the templatized context to convert more human-like narrative. In total, we have created three different prompts each for PL[1D], FOL[2D], and NM[3D. The prompts are designed to ensure that the logical connection is established in the narrative. Each rule of PL, FOL, and NM has a unique logical progression that should be followed in the narrative is the motivation for us to go with three different instruction-based prompts. In the prompt, used for PL rules, the logical rule is also mentioned in two parts "Condition" and "Situation". Consider the example of logical rule MT - "If p then q; not q; therefore not p", the templatized context will have "If p then q; not q;" so condition will be the first part "If p then q" while the situation will be the second part "not q". In other terms, the specific conditions are the rules to be followed while the situation is the case present in the logical rule. In contrast, for FOL, we do not need to focus more on the specific condition rather we have to make sure a generalized case is present in the narrative with 1-2 specific sentences related to the rule. For NM, we do not have such a rule instead we have a logical connection between sentences and hence we go with the instruction-based prompting.

E Task Instance Generation

This section discusses the task instance generation step, which is the last step in Fig. 2 in detail. The overall section is comprised of two subsections for BQA and MCQA. The subsection related to BQA

discusses the variation generation step in detail with question generation while the MCQA subsection provides details of prompts used for incorrect option generation along with five pre-defined sets of questions.

E.1 Task 1- BQA (variation generation)

As seen in Fig. 2, BQA has a narrative along with question, and answer pairs with variation in question and corresponding answer. As discussed in the main paper Data creation3.2, the narrative is created using the pre-defined rules for PL, FOL, and NM while a question is asked based on what can be entailed from the given context. For example, HS can be defined as "If p then q; if q then r; therefore, if p then r" and the narrative will have "If p then q; if q then r" and the question is asked as "Can we conclude if p then r?" having answer as "yes". Now, other variations in the question are asked in the following ways by negating the sentence (p and r) comb: Variation 1: "Can we conclude if ¬p then r?", Variation 2: "Can we conclude if p then ¬r?", Variation 3: "Can we conclude if ¬p then ¬r?". If there's only one axiom (p) present in the question then there are only a 2 variations that can be made asking about (p) and (¬p).

E.2 Task 2-MCQA (Question selection and incorrect option generation)

We have MCQA as task 2 in LogicBench. In this, we have one correct option from the four options, and three options are incorrectly generated using prompting. The question is a bit different from the BQA question formation as we have MCQA and have to identify which conclusion can be derived from the narrative. The question is randomly selected from the pre-defined set of questions. Here, the five different questions are as follows:

1. What would be the most appropriate conclusion based on the given context?

2. Considering the provided context, what conclusion would be deemed most suitable?

3. In light of the context provided, what conclusion can be considered the most appropriate?

4. Based on the context, what conclusion would be deemed most suitable?

5. Taking into account the context provided, what conclusion would be most appropriate?

The narrative for MCQA is the same as the BQA's narrative and there's no change. Comparing BQA methodology with MCQA, MCQA's correct option is the question asked in the BQA which can be concluded from the logical rule present in the narrative. For example, HS can be defined as "If p then q; if q then r; therefore, if p then r" and the narrative will have "If p then q; if q then r" and the question is randomly chosen from the set of questions as mentioned above and the correct option from the different multiple choice will be "if p then r". Based on the information present, we have generated an incorrect option using prompting[4E.2]. Refer this link for more examples - https://anonymous.4open.science/r/LogicBench-EEBB.

4. Prompt for incorrect option generation:

> **Context:** Narrative
> **Question:** Randomly chosen question
> **Correct option:** Conclusion from narrative
>
> Create three more irrelevant, incorrect options to create multiple choice. Sometime use different persons name as well in creating incorrect options.
> **Instruction to generate options:**
> 1. Generate only three unique choices each in new line.
> 2. Do not use any prefix while generating.

F Mitigation of Errors in *LogicBench*

While validating *LogicBench(Eval)*, we encountered errors within the synthetically generated narrative-based context. We mitigate these errors manually, categorizing them into two groups: (i) eliminating leakage of logical conclusions and (ii) ensuring the inclusion of intended logical premises. In the first category, we found $\sim 15\%$ of the narrative-based contexts (out of 500 total instances) were found to explicitly present the logical conclusion as a response to the question, bypassing the logical derivation process. This enables the model to extract the final logical conclusion from the context rather than derive it logically. In the second category, we found $\sim 8\%$ of the narrative-based contexts (out of 500 total instances) where narration lacked some necessary premise sentences crucial for reaching a logical conclusion. To address this issue, we manually incorporated those sentences to ensure the quality of the data. For the MCQA task, during the generation of three incor-

rect options, we found instances where the model produces two semantically similar options, resulting in the creation of ambiguous choices in $\sim 11\%$ of the cases (out of a total of 500 instances). We manually mitigated all these errors from the data instances ensuring the high quality of our validation data. We believe that these two versions aim to accommodate different evaluations to explore the logical reasoning capabilities of LLMs.

G Few-shot Experiments

This section discusses the performance of the different LLMs in a few-shot setting on the *LogicBench(Eval)*$_{BQA}$. Here, we only present a case study on the BQA task. For the fair comparison with Table 6, we analyze an average performance across $A(Yes)$. Table 14 shows the performance for each inference rule and reasoning patterns achieved by Llama-2, Mistral, Gemini, ChatGPT, and GPT-4.

As suggested in (Lu et al., 2022), prompting models are sensitive to in-context examples. Hence, we see mixed performance in Table 14 across all models. From Table 14, we can observe that in-context examples are helpful for Llama-2 since it consistently outperforms zero-shot baselines by large margins in terms of $A(Yes)$. Llama-2 is remarkably good at following the in-context exemplars and mimicking the process to reach the correct conclusions. Thus, leveraging the in-context exemplars, Llama-2 achieves high accuracy in a few-shot setting. Specifically, Mistral consistently shows degraded performance for all logic types. However, ChatGPT improves performance on NM reasoning. Improved performance in NM reasoning demonstrates that the inclusion of in-context examples enhances the ability of these models to comprehend the nuanced meanings of logical terms such as "usually" and "typically". In particular, we see that Gemini and GPT-4 improve performance on PL and FOL, respectively, but show competitive performance on NM.

H Experimental Setup

H.1 Extended Discussion on Experiments

Zero-shot setting We evaluate GPT-4, and ChatGPT (GPT-3.5-Turbo) by utilizing their APIs provided by OpenAI[3]. We evaluate Google Gemini-Pro by utilizing its API provided by Google[4] The evaluation is conducted on the versions of GPT-4, ChatGPT, and Gemini released in January 2024. It's important to note that these models are regularly updated, so when reproducing the results presented in Table 6 and Table 7 (main paper), there is a possibility of variations. For Llama-2 and Mistral, we utilize the 7B-Chat, and 7B-Instruct-v0.2 versions, respectively, from the huggingface model repository[5].

Experiments on other logic datasets In single and multi-task experiments on other logic datasets, we fine-tune the T5-large model for 10 epochs with a batch size of 16, 1024 maximum input length, an adaptive learning rate of $5e-05$, and an AdamW optimizer for each experiment. All experiments are performed using NVIDIA RTX A6000 GPUs.

H.2 Prompts

All the experiments conducted in the zero-shot setting were performed using three distinct prompts. The reported results in Table 6 (main paper) represent the average performance across these prompts. All the prompts follow the common pattern which includes task description and formatting instructions. The following are the three different prompts utilized in the experiments:

Prompt 1:

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

Prompt 2:

> Let's think step-by-step to answer the question given context. Answer the question ONLY in 'yes' or 'no'. Please use the below format:
> **Context**: [text with logical rules]
> **Question**: [question based on context]
> **Reasoning steps**: [generate step-by-step reasoning]
> **Answer**: Yes/No

[3] https://platform.openai.com/docs/guides/gpt
[4] https://ai.google.dev/
[5] https://huggingface.co/models

Prompt 3:

> 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 based on context]
> **Reasoning steps:** [generate step-by-step reasoning]
> **Answer:** Yes/No

For MCQA task, the reported results in Table 7 (main paper) represent the average performance across three prompts. All the prompts follow the common pattern which includes task description and formatting instructions. The following are the three different prompts utilized in the experiments:

Prompt 1:

> Given the context that contains rules of logical reasoning in natural language, question, and options, perform step-by-step reasoning to answer the question. Answer the question ONLY in 'choice_1' or 'choice_2' or 'choice_3' or 'choice_4'. DO NOT generate any text with it. Based on context and reasoning steps, choose the correct option.
> Please use the below format:
> **Context:** [text with logical rules]
> **Question:** [question based on context]
> **Options:** [four different choices]
> **Reasoning steps:** [generate step-by-step reasoning]
> **Answer:** choice_1 / choice_2 / choice_3 / choice_4

Prompt 2:

> Let's think step-by-step to choose the correct option for the question given context. Answer the question ONLY in 'choice_1' or 'choice_2' or 'choice_3' or 'choice_4'. DO NOT generate any text with it. Based on context and reasoning steps, choose the correct option.
> Please use the below format:
> **Context:** [text with logical rules]
> **Question:** [question based on context]
> **Options:** [four different choices]
> **Reasoning steps:** [generate step-by-step reasoning]
> **Answer:** choice_1 / choice_2 / choice_3 / choice_4

Prompt 3:

> Given the context, question, and options, think step-by-step logically to answer the question by choosing the correct option. Answer the question ONLY in 'choice_1' or 'choice_2' or 'choice_3' or 'choice_4'. DO NOT generate any text with it. Based on context and reasoning steps, choose the correct option.
> Please use the below format:
> **Context:** [text with logical rules]
> **Question:** [question based on context]
> **Options:** [four different choices]
> **Reasoning steps:** [generate step-by-step reasoning]
> **Answer:** choice_1 / choice_2 / choice_3 / choice_4

I Further Discussion on Results

Effect on other logic datasets Our experiments were carried out in two settings: single-task (fine-tuning and evaluation on one dataset) and multi-task (fine-tuning on all four datasets combined, with separate evaluations for each dataset). App. H describes a detailed experimental setup.

Table 11 represents the accuracy comparison between LogicT5 and baseline T5-large in both single-task and multi-task settings.

The results indicate that training LLMs on *LogicBench(Aug)* has a greater impact on logic datasets that primarily focus on logical reasoning, such as FOLIO and LogicNLI. Hence, we can observe that LogicT5 consistently outperforms the baseline for LogicT5 and FOLIO. However, LogiQA and ReClor encompass other forms of reasoning in addition to logical reasoning, hence, LogicT5 demonstrates competitive performance on them. More analysis is presented in Appendix I.

PL *vs.* FOL To get an indication of this we gave GPT-4 prompt - "Give twenty statements that have 'if' and 'then' in them". From the results, we can

Methods	Models	LogiQA	FOLIO	LogicNLI	ReClor
Single-Task	T5-large	16.8	69.6	82.3	35.4
	LogicT5	**16.9**	**71.2**	**84.4**	**36.8**
Multi-Task	T5-large	**21.8**	83.8	68.2	**42.8**
	LogicT5	19.7	**85.6**	**69.8**	40.0

Table 11: Performance comparison between LogicT5 and baseline T5-large in terms of accuracy.

observe that 13 of the 20 sentences in response to the prompt were FOL kind and only 7 were propositional kind. This shows that LLMs' comprehend simple FOL sentences, thus, showing high overall accuracy on simpler FOL compared to PL. However, we interestingly observed that, as the depth increases, PL-based examples carrying a specific case/story become easier to understand compared to FOL-based generalized context.

Negations are hard to understand when embedded with logical rules. Here, Table 15 presents reasoning chains corresponding to PL and FOL supporting the effect of negation on their predictions.

Lower performance of NM compared to PL for MCQA. We present reasoning chains corresponding to PL and NM supporting their behavior on both BQA and MCQA tasks at https://anonymous.4open.science/r/LogicBench-EEBB.

Rule	Generate Sentences in Step 1	NL logical expressions
MT	p: Liam finished his work early. ∼p: Liam did not finish his work early. q: He will order pizza for dinner. ∼q: He will not order pizza for dinner.	**Context:** Liam had finished his work early for the day, which meant that he would typically have ordered pizza for dinner. However, on this particular day, he decided against ordering pizza and opted for something else instead. **Question:** Does this imply that liam didn't finish his work early?
BD	p(x): someone drinks lots of water q(x): they will feel hydrated r(x): they eat too much sugar s(x): they will experience a sugar crash p(a): Jane drinks lots of water ∼p(a): Jane does not drink lots of water q(a): she will feel hydrated ∼q(a): she will not feel hydrated r(a): she eats too much sugar ∼r(a): she does not eat too much sugar s(a): she will experience a sugar crash ∼s(a): she will not experience a sugar crash	**Context:** If an individual consumes a significant amount of water, they will experience a state of hydration. Conversely, if excessive amounts of sugar are ingested, a sugar crash will ensue. It is known that at least one of the following statements is true: either the Jane consumes ample water or she will not experience a sugar crash. However, the actual veracity of either statement remains ambiguous, as it could be the case that only the first statement is true, only the second statement is true, or both statements are true. **Question:** Can we say at least one of the following must always be true? (a) she will feel hydrated and (b) she doesn't eat too much sugar
MP	p(x): someone is exhausted. q(x): they will take a rest. p(a): Jack is exhausted. ∼p(a): Jack is not exhausted. q(a): he will take a rest. ∼q(a): he will not take a rest.	**Context:** If someone is extremely tired, then they will seek some rest and relaxation. Today, Jack finds himself utterly exhausted. **Question:** Does this entail that he will take rest?
DS	p: Levi is not studying for his exams ∼p: Levi is studying for his exams q: Maya is writing a book ∼q: Maya is not writing a book	**Context:** Either Levi is not studying for his exams or Maya is writing a book, or maybe both. It was unclear which choice to make as he didn't know if either of the options was true. However, it turned out that Levi decided to prioritize his exams and focus on studying. He knew that in order to succeed, he needed to dedicate his time and energy to preparing for the upcoming tests. **Question:** Does this mean that Maya is writing a book?
HS	p(x): someone buys all the necessary supplies. q(x): they can start the project. r(x): they will finish it on time. p(a): Lily bought all the necessary supplies. ∼p(a): Lily did not buy all the necessary supplies. q(a): She can start the project. ∼q(a): She can not start the project. s(a): She will finish it on time. ∼s(a): She will not finish it on time.	**Context:** If all the necessary supplies have been purchased by someone, then they can initiate the project. Once the project is started by someone, they will complete it within the expected timeframe. **Question:** If lily bought all the necessary supplies, does this mean that she will finish it on time?
CD	p: Harry goes to the park. ∼p: Harry does not go to the park. q: he will have a picnic with his family ∼q: he will not have a picnic with his family r: he goes to the beach ∼r: he does not go to the beach s: he will swim in the ocean ∼s: he will not swim in the ocean	**Context:** Harry had two options for his day off: going to the park or going to the beach. If he decided to go to the park, it meant that he would have a lovely picnic with his family. On the other hand, if he chose the beach, he would be able to enjoy a refreshing swim in the ocean. The only thing that was certain was that either Harry would go to the park or he would go to the beach. It was unclear which option he would ultimately choose, as it could be that he would go to the park alone, go to the beach alone, or even do both activities. **Question:** Can we say at least one of the following must always be true? (a) he will have a picnic with his family and (b) he will swim in the ocean
DD	p: I order takeout ∼p: I did not order takeout q: I will save time ∼q: I will not save time r: I cook a meal ∼r: I did not cook a meal s: I will save money ∼s: I will not save money	**Context:** If I decide to order takeout, it means I will save time. On the other hand, if I choose to cook a meal, it means I will save money. The interesting thing is that I am in a situation where I won't be able to save time or money. It is uncertain whether I won't save time or I won't save money, or it could even be both scenarios. The only thing that is clear is that at least one of these possibilities is true. **Question:** Can we say at least one of the following must always be true? (a) i don't order takeout and (b) i don't cook a meal
CT	p: Tom is an avid reader ∼p: Tom is not an avid reader q: he devours books of all genres ∼q: he does not devour books of all genres	**Context:** At least one of two things is true about Tom - he is either an avid reader or he devours books of all genres. We are unsure which one of these statements is true or if both are true. It could be that only the first statement is true, or only the second statement is true, or even that both are true. **Question:** Can we say at least one of the following must always be true? (a) he devours books of all genres or (b) tom is an avid reader
MI	p: he is not eating healthy ∼p: he is eating healthy q: he will not gain weight ∼q: he will gain weight	**Context:** Once upon a time, there was a guy named John who was very conscious about his weight. He knew that if he didn't eat healthy, he wouldn't be able to maintain his desired weight. John understood the condition that if he didn't eat healthy, he wouldn't be able to gain weight. **Question:** Based on context, can we say, at least one of the following must always be true? (a) he is eating healthy and (b) he will not gain weight
EG	p(x): someone has coding skills a: Sheila p(a): Sheila is a proficient programmer	**Context:** Sheila possesses exceptional coding skills and is highly proficient in programming. She has mastered various programming languages and is known for her ability to write efficient and error-free code. Sheila's expertise in programming allows her to solve complex problems with ease. Her colleagues often seek her assistance when facing technical challenges as she consistently delivers high-quality solutions. **Question:** does this mean that someone has coding skills?
UI	p(x): students need to take an exam to complete their degree a: Reema p(a): Reema is a student.	**Context:** All students are required to take an examination in order to fulfill the requirements for their degree. Reema, being a student, is also expected to fulfill the requirements. **Question:** Does reema need to take an exam to complete her degree?

Table 12: Illustrative examples of NL logical context and questions created using sentences that are generated in §3.2.1 for rules covered in PL and FOL for BQA.

Rule	Generate Sentences in Step 1	NL logical expressions
DRI	p: Cats and dogs are mammals. q: Mammals typically have fur. r: Cats don't have fur. s: Dogs are loyal animals. t: Dogs have fur. ∼t: Dogs don't have fur.	**Context:** Once upon a time, in a land filled with animals, there were two popular mammalian creatures, cats and dogs. Mammals typically possessed a coat of fur, which kept them warm and protected. However, cats were an exception to this rule, as their bodies lacked fur. Nonetheless, both cats and dogs were beloved by many for their unique traits. Dogs, known for their loyalty, were particularly cherished by humans. **Question:** Does this imply that dogs have fur?
DRS	p: John and Mary are parents. q: Parents are usually loving and supportive. r: Parents are normally responsible. s: Mary isn't loving and supportive. t: John is not responsible. u: Mary is responsible. ∼u: Mary isn't responsible. v: John is loving and supportive. ∼v: John isn't loving and supportive.	**Context:** John and Mary were expecting their first child, filled with the anticipation and excitement that all parents feel. Parents are usually loving and supportive. Parents are normally responsible. However, something seemed amiss in their relationship. Mary, usually affectionate and caring, seemed distant and uninvolved. On the other hand, John, known for his responsible nature, started neglecting his duties and became unreliable. **Question:** Does this imply that Mary is responsible and John is loving and supportive?
DRD	p: Jenny and Anna are tall. q: Tall people usually play basketball. r: Anna is possibly an exception to this rule. s: Jenny plays basketball. ∼s: Jenny doesn't play basketball.	**Context:** Jenny and Anna are known for their tall stature, which is often associated with playing basketball. However, Anna might be an exception to this norm. **Question:** Does this entail that Jenny plays basketball?
DRO	p: Hummingbirds are birds. q: Birds migrate south for the winter. r: Hummingbirds do not migrate south for the winter. s: All other birds than hummingbirds migrate south for the winter. ∼s: All other birds than hummingbirds don't migrate south for the winter.	**Context:** In the bird kingdom, there are many different species that possess unique characteristics. One such species is the hummingbird, known for its ability to hover in mid-air and its vibrant colors. While most birds engage in the annual migration south for the winter, the hummingbird chooses to stay put and brave the cold weather. This decision sets the hummingbird apart from its fellow avian companions, as it relies on its resilience and resourcefulness to survive the harsh conditions. **Question:** Does this mean that all other birds than hummingbirds migrate south for the winter?
RE1	p: Cats, dogs, and horses are animals. q: Animals are usually considered to be intelligent creatures. r: At least one of the cats or dogs is not considered intelligent. s: Horses are considered to be intelligent creatures. ∼s: Horses aren't considered to be intelligent creatures. t: Exactly one of the cats or dogs is not considered intelligent. ∼t: Exactly one of the cats or dogs is considered intelligent.	**Context:** In a world where animals are often regarded as intelligent creatures, there is a captivating tale that revolves around cats, dogs, and horses. It is commonly believed that most animals possess a level of intellect. However, there is an intriguing twist to this belief as it is known that either cats or dogs are not considered particularly intelligent. As the story unfolds, we delve into the lives of these remarkable creatures, their interactions, and the unique qualities that each of them possesses. **Question:** Does this entail that horses are considered to be intelligent creatures and exactly one of the cats or dogs is not considered intelligent?
RE2	p: cats normally meow. q: At least one species of cat doesn't meow. r: Exactly one species of cat doesn't meow. ∼r: Exactly one species of cat meows.	**Context:** In the realm of cat communication, meowing serves as a fundamental aspect of their vocal repertoire. However, intriguingly enough, there exists a distinct species of cat that deviates from this conventional norm. This peculiar feline defies the expectations associated with its kind by refraining from emitting any meows whatsoever. **Question:** Does this entail that exactly one species of cat doesn't meow?
RE3	p: Cars have four wheels. q: wheels normally have spokes. r: at least one wheel does not have spokes. s: Cars have four wheels with spokes. ∼s: Cars don't have four wheels with spokes.	**Context:** In a world where cars were known for having four wheels, it was considered a common fact that wheels typically came equipped with spokes. However, amidst this widespread understanding, there was an exception. At least one wheel defied this norm and stood out from the rest by not having any spokes at all. **Question:** Does this imply that cars have four wheels with spokes?
RAP	p: John asserts that Sally was in the store. q: Jane asserts that Sally was not in the store. r: John's evidence is more reliable than Jane's. ∼r: John's evidence is less reliable than Jane's. s: Sally was in the store. ∼s: Sally wasn't in the store.	**Context:** In the midst of a heated argument, John adamantly claims that Sally was present at the store. However, Jane strongly opposes John's assertion, insisting that Sally was indeed absent from the store. **Question:** If John's evidence is more reliable than Jane's, does this mean that Sally was in the store?

Table 13: Illustrative examples of NL logical context and questions created using sentences that are generated in §3.2.1 for NM logic for BQA.

Type	Rule	Llama-2		Mistral		Gemini		ChatGPT		GPT-4	
		$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$	$A(No)$	$A(Yes)$
PL	HS	100.0	25.3	74.3	16.7	83.3	26.9	84.2	27.9	93.8	53.1
	DS	66.7	26.9	76.6	66.7	95.2	32.8	87.5	33.3	92.2	93.8
	CD	100.0	55.6	97.1	42.2	98.1	73.1	100.0	57.1	100.0	80.0
	DD	93.8	79.2	61.1	59.1	94.7	90.0	64.3	83.3	64.3	83.3
	BD	76.9	25.4	76.4	37.5	77.8	24.6	82.4	30.4	85.1	76.9
	MT	100.0	25.6	89.3	32.7	92.9	28.1	100.0	30.3	91.5	48.5
	MI	55.6	51.6	57.9	57.1	66.7	66.7	64.0	73.3	90.5	94.7
	CT	84.7	52.6	97.4	45.2	97.4	48.7	85.7	36.8	96.7	90.0
	Avg	**84.7**	**42.8**	**78.8**	**44.7**	**88.3**	**48.9**	**83.5**	**46.6**	**89.2**	**77.5**
FOL	EG	90.9	27.5	76.8	36.4	88.9	27.6	77.4	29.6	90.4	53.6
	UI	75.0	25.8	79.7	83.3	88.0	31.5	85.7	33.3	92.1	88.2
	MP	100.0	69.0	100.0	64.5	95.2	100.0	100.0	100.0	100.0	95.2
	HS	100.0	100.0	100.0	83.3	100.0	100.0	100.0	100.0	100.0	100.0
	DS	100.0	51.3	97.3	44.2	87.5	65.0	100.0	50.0	92.9	66.7
	CD	90.5	94.7	61.9	63.2	81.0	84.2	81.0	84.2	74.1	100.0
	DD	80.0	25.3	74.7	0.0	88.0	30.9	78.3	29.4	85.1	76.9
	BD	72.7	58.6	77.8	58.1	81.8	94.1	79.2	93.8	90.5	94.7
	MT	85.7	69.2	57.7	64.3	82.4	73.9	86.4	94.4	66.7	100.0
	Avg	**88.3**	**57.9**	**80.7**	**55.2**	**88.1**	**67.5**	**87.5**	**68.3**	**88.0**	**86.2**
NM	DRI	86.2	70.6	62.5	62.5	80.4	90.9	74.0	90.0	95.2	100.0
	DRS	47.4	47.6	77.3	83.3	61.9	70.6	60.7	75.0	90.0	90.0
	DRD	100.0	66.7	100.0	62.5	78.6	68.0	63.6	55.2	70.4	92.3
	DRO	58.1	14.3	66.7	7.7	62.8	9.1	64.1	14.6	76.5	33.3
	RE1	60.9	64.7	58.6	72.7	66.7	85.7	60.6	100.0	87.0	100.0
	RE2	0.0	25.0	88.0	30.9	84.8	32.6	77.8	25.8	86.5	46.4
	RE3	52.6	52.4	68.0	80.0	82.4	71.4	62.1	81.8	61.5	71.4
	RAP	50.0	50.0	65.5	90.9	62.5	100.0	55.6	100.0	87.0	100.0
	Avg	**56.9**	**48.9**	**73.3**	**61.3**	**72.5**	**66.0**	**64.8**	**67.8**	**81.8**	**79.2**

Table 14: Performance of LLMs in few-shot setting in terms of label-wise accuracy on LogicBench(Eval), where $A(Yes)$ and $A(No)$ denote the accuracy for the Yes and No labels, respectively.

Rule	Context/Question	Actual/Pred	Reasoning Chains
PL(MT)	Context: if elijah has a new job, then he will get a raise. however, he would not be receiving the raise he had anticipated. Question: does this infer that elijah doesn't have a new job?	Yes/No	1. The context provides a conditional statement: If Elijah has a new job, then he will get a raise. 2. The context also states that Elijah would not be receiving the raise he had anticipated. 3. The second statement does not negate the first; it only states that the raise is not as expected. 4. The lack of the anticipated raise does not necessarily mean Elijah doesn't have a new job; it could mean that the raise with the new job is different from what he expected. 5. Therefore, the information provided does not allow us to infer that Elijah doesn't have a new job based solely on the raise situation.
PL(MT)	Context: james had always been forgetful, especially when it came to wearing a sweater. his mother had repeatedly told him, 'if he forgets to wear a sweater, then he will catch a cold.' throughout the day, he went about his activities, oblivious to the dropping temperature. hours later, he returned home, but james didn't catch a cold. Question: does this infer that james didn't forget to wear a sweater?	Yes/No	1. The mother's rule establishes a conditional relationship: If James forgets to wear a sweater (condition), then he will catch a cold (result). 2. It is stated that James did not catch a cold upon returning home. 3. According to the mother's rule, if the result (catching a cold) did not occur, we cannot conclusively determine whether the condition (forgetting to wear a sweater) happened or not. 4. The absence of the result (catching a cold) does not necessarily mean the condition (forgetting to wear a sweater) did not occur, as there could be other factors at play that prevented James from catching a cold. 5. Therefore, from the information given, we cannot infer with certainty that James remembered to wear a sweater.
FOL(DD)	Context: if an individual operates a vehicle with caution, they will prevent any accidents from occurring. however, if they decide to accelerate and exceed the speed limit, they will be issued a fine. there is a certainty that at least one of the following statements is true: either john will not be able to avoid an accident or he will not face a penalty. it is unclear which of these statements is accurate, as it could be only one of them, or it is possible that both of them hold true. Question: can we say at least one of the following must always be true? (a) john doesn't drive carefully and (b) he doesn't speed up	Yes/No	1. According to the context, if an individual operates a vehicle with caution, they will prevent any accidents from occurring. 2. Also, if they decide to accelerate and exceed the speed limit, they will be issued a fine. 3. It is stated that there is a certainty that at least one of the following statements is true: either John will not be able to avoid an accident or he will not face a penalty. 4. It is unclear which of these statements is accurate, as it could be only one of them, or it is possible that both of them hold true. 5. The question asks if we can say at least one of the following must always be true: (a) John doesn't drive carefully and (b) he doesn't speed up. 6. Based on the given information, we cannot say that at least one of the statements (a) and (b) must always be true.
FOL(DD)	Context: if someone is taking care of thier health, then they will be fit. however, indulging in unhealthy habits can make individuals susceptible to various diseases. the truth is, we can be certain about at least one of the following possibilities: either jenny won't be fit and healthy, or she won't be prone to diseases. it is important to note that we are unaware of which statement is accurate. it could be the case that only the first statement is true, only the second statement is true, or both statements hold validity. Question: can we say at least one of the following must always be true? (a) jenny takes care of her health and (b) she doesn't indulge in unhealthy habits	No/Yes	1. According to the context, if someone takes care of their health, they will be fit. 2. Indulging in unhealthy habits can make individuals susceptible to various diseases. 3. The context states that we can be certain about at least one of the following possibilities: either Jenny won't be fit and healthy or she won't be prone to diseases. 4. It is important to note that we are unaware of which statement is accurate. 5. The possibilities are: a) Jenny takes care of her health and is fit. b) Jenny doesn't indulge in unhealthy habits and is not prone to diseases. c) Both statements hold validity. 6. Therefore, at least one of the following must always be true: (a) Jenny takes care of her health and (b) she doesn't indulge in unhealthy habits.

Table 15: Case study of reasoning chains for PL(MT) generated from ChatGPT, and FOL(DD) generated from GPT-4.