$A \wedge B \Leftrightarrow B \wedge A$: Evaluating and Improving the Logical Reasoning Ability of Large Language Models

Anonymous ACL submission

Abstract

We introduce LogicAsker, a novel approach for evaluating and enhancing the logical reasoning capabilities of large language models (LLMs) such as ChatGPT and GPT-4. Despite their prowess in tasks like writing assistance, code generation, and machine translation, assessing LLMs' ability to reason has been challenging. Traditional evaluations often prioritize accuracy on downstream tasks over direct assessments of reasoning processes. LogicAsker addresses this gap by employing a set of atomic reasoning skills grounded in propositional and predicate logic to systematically examine and improve the reasoning prowess of LLMs. Our methodology reveals significant gaps in LLMs' learning of logical rules, with identified reasoning failures ranging from 29% to 90% across different models. Moreover, we leverage these findings to construct targeted demonstration examples and fine-tune data, notably enhancing logical reasoning in models like GPT-4o by up to 5%. To our knowledge, this is the first effort to utilize test case outcomes to effectively refine LLMs' formal reasoning capabilities. We will make our code, data, and results publicly available to facilitate further research and replication of our findings.

1 Introduction

Large language models (LLMs), such as OpenAI's GPT series have significantly impacted natural language processing, excelling in a variety of tasks including text generation, machine translation, and code generation (Gao et al., 2022, 2023a; Jiao et al., 2023).

Reasoning, defined as the cognitive process of using logic to draw conclusions from given facts (Wei et al., 2022b,a), is crucial for complex interactions that go beyond text generation. Accurately assessing this ability in LLMs is essential, yet challenging, as models may correctly perform tasks merely relying on shortcuts such as pattern recognition without truly engaging in logical reasoning (Huang and Chang, 2022; Huang et al., 2023; Liu et al., 2023). Consider the following inference example: `Either it is raining, or Tom will play football; if it rains, then the floor will be wet; the floor is dry; therefore, Tom will play football.` We may encounter the following challenges: 1) It's unclear if a correct LLM response is due to reasoning or simple heuristics like word correlations (e.g., "dry floor" is more likely to correlate with "playing football"). 2) If an LLM fails, pinpointing the specific breakdown in reasoning is difficult (i.e., inferring not raining from the floor being dry or inferring playing football from not raining). 3) Current systems lack comprehensive test cases that encompass various formal reasoning types beyond implication, such as logical equivalence (e.g., A and B are true; therefore, B and A are true. 4) Evaluating an LLM's reasoning on such cases offers limited insight into enhancing its reasoning capabilities.

To better handle these challenges, a well-performing testing framework should be able to define a set of skills that a) directly correspond to the reasoning process, b) cannot be further divided, c) cover all formal logical reasoning scenarios, and d) can identify LLMs' weaknesses and facilitate improving LLMs' performance. Property a) ensures that the task cannot be accomplished by other approaches, such as inferring from the correlations of words, and the evaluation result directly reflects the model's reasoning ability. Property b) and c) ensure that the set of skills is fundamental and comprehensive, which can provide helpful insights to accomplish Property d).

We introduce LogicAsker, an automatic framework designed to evaluate and enhance LLMs' formal reasoning skills using Minimum Functionality Tests (MFTs) (Ribeiro et al., 2020), akin to software engineering's unit tests, which utilize straightforward examples to assess specific behav-

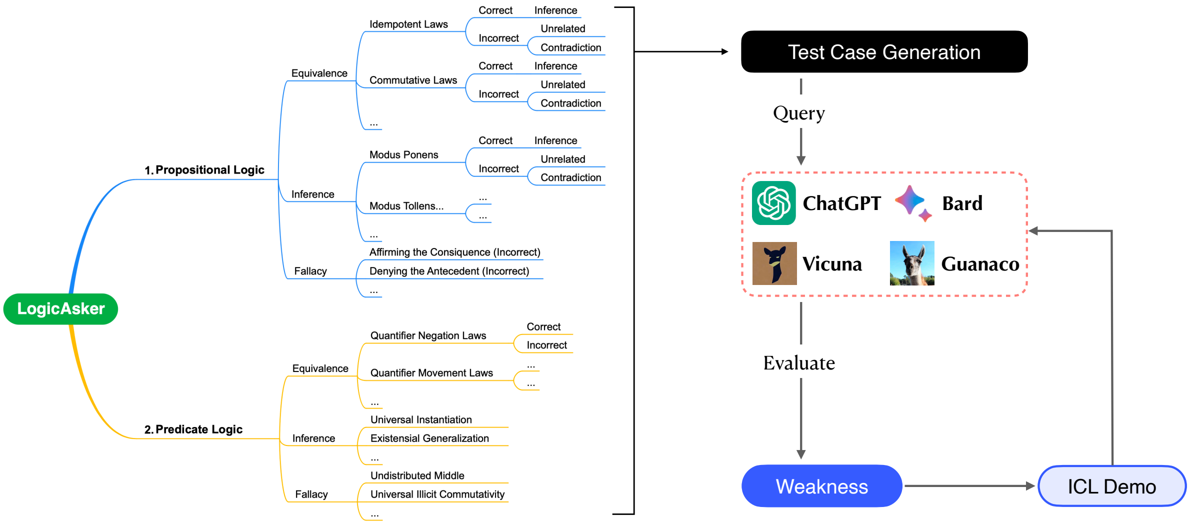

Figure 1: Overview of the LogicAsker framework.

iors. These tests help identify when models rely on shortcuts rather than genuinely mastering a skill (Ribeiro et al., 2020). Specifically, LogicAsker builds a set of atomic skills from foundational principles of propositional and predicate logic, two fundamental systems used to formalize reasoning procedures (Partee et al., 1990), together with common logical fallacies (Hurley and Watson, 2020). Based on the skill set, LogicAsker generates reasoning questions by translating standard logic expressions into natural language, assesses LLMs' accuracy per skill, pinpoints weaknesses, and creates in-context-learning (Brown et al., 2020) examples and fine-tuning data to bolster reasoning abilities. In addition, for each skill, LogicAsker uses diverse vocabulary to frame various natural language queries, computing average performance to minimize biases from word correlations.

Table 1 demonstrates that LogicAsker complements existing frameworks by providing a comprehensive evaluation scope and utilizing outcomes to enhance LLMs' reasoning capabilities, while other datasets often face data leakage and are scope-limited. LogicAsker serves as an extensive diagnostic tool for LLMs' formal reasoning, significantly exceeding the coverage of comparable tools and enabling detailed assessments across diverse reasoning rules such as inferences, quantifiers, and fallacies. Scaling up the scope presents significant challenges due to the complexity of designing algorithms capable of processing various logical rules and translating them into natural language. Despite these complexities, LogicAsker uniquely integrates all formal logical rules and common fallacies, facilitating robust testing and refinement of reasoning capabilities.

We evaluated LogicAsker's performance through extensive testing on six state-of-the-art (SOTA) LLMs (Hugging Face, 2024), including four close-source LLMs (GPT-4o, GPT-4, ChatGPT, and Gemini-1.5) and two open-source LLMs (Llama3 and Mixtral). Our findings reveal that LogicAsker's test cases effectively pinpoint logical reasoning failures across these models, with error rates (i.e., $1 -$ accuracy) between 29% and 90%. These test cases also facilitate the creation of in-context learning examples and fine-tuning data, thereby enhancing logical reasoning capabilities. For instance, applying LogicAsker's cases to GPT-4o improved its reasoning accuracy from 92% to 97%. All resources will be available for reproduction and further research[1].

We summarize the main contributions of this work as follows:

- We are the first work that formally defines a comprehensive set of 30 atomic and 208 extended skills necessary for LLMs to execute formal reasoning based on propositional and predicate logic.
- We develop LogicAsker, a fully automatic tool that utilizes atomic skills to generate test cases to assess and enhance LLMs' reasoning abilities, marking a first in utilizing test results to directly improve LLM performance.
- We conduct a thorough empirical evaluation of the logical reasoning abilities of six SOTA LLMs.

[1] https://drive.google.com/drive/folders/19xj5XjnSbtlY1vvT0kbcKfY1FfvCnE9j?usp=share_link

We demonstrate that the test results by LogicAsker can be used to effectively evaluate and improve the performance of LLMs.

2 Preliminaries

2.1 Formal Analysis of Reasoning Abilities

"Reasoning" can be characterized into formal reasoning and informal reasoning. The former is a systematic and logical process that follows a set of rules and principles, and the reasoning within these systems will provide valid results as long as one follows the defined rules (e.g., all A are B, all B are C; therefore, all A are C). The latter is a less structured approach that relies on intuition, experience, and common sense to draw conclusions and solve problems (Huang and Chang, 2022; Bronkhorst et al., 2020) (e.g., Hong Kong residents have a high life expectancy; this is probably because they have healthy living habits). Generally, formal reasoning is more structured and reliable, whereas informal reasoning is more adaptable and open-ended but may be less reliable. In this paper, we focus on the formal reasoning process to systematically analyze LLMs' reasoning abilities.

To formalize reasoning procedures, two fundamental systems are usually adopted, namely, propositional logic and predicate logic. The former one deals with propositions or statements that can be either true or false, and utilizes logical operators including $\wedge$ (and), $\vee$ (or), $\neg$ (not), $\rightarrow$ (inference), and $\leftrightarrow$ (bidirectional) to connect these statements. The latter one, in contrast, extends propositional logic to deal with more complex statements that involve variables, quantifiers, and predicates. Both propositional logic and predicate logic contain various rules for the reasoning process. These rules can be categorized into equivalence rules and inference rules. Equivalent rules summarize the basic expressions that are equivalent in terms of truth value (e.g., $\neg(P \wedge Q) \Leftrightarrow (\neg P) \vee (\neg Q)$). Inference rules summarize the basic valid inference rules (e.g., from the premises: $A \rightarrow B$, and A, we can infer B).

We refer to (Partee et al., 1990) for a more detailed explanation. Table 7-9 in Appendix A list common inference rules in predicate logic and propositional logic. Besides inference rules, formal logic systems can also express common logical fallacies, i.e., arguments that may sound convincing but are based on faulty logic and are, therefore, invalid. We list the common logical fallacies in Table 10.

2.2 Minimum Functionality Test

In this paper, we adopted the concept of Minimum Functionality Tests (MFTs), introduced in (Ribeiro et al., 2020), to evaluate the reasoning ability of LLMs. MFTs are analogous to unit tests in software engineering, where a collection of simple examples is used to check a specific behavior within a capability. These tests involve creating small and focused datasets that are particularly effective in detecting whether models resort to shortcuts to handle complex inputs, rather than truly mastering the capability.

To apply MFTs in evaluating the reasoning ability of LLMs, we treated each formal logical rule as an independent task and generated abundant test cases for each task. Each test case was designed to trigger logical failures in the LLMs, allowing us to assess the strengths and weaknesses of LLMs in the logical reasoning process, and providing a solid foundation for further analysis and improvement.

3 LogicAsker

In this section, we introduce the design and implementation of LogicAsker, a novel tool to trigger logical reasoning failures in large language models. Figure 1 overviews the workflow of LogicAsker, which consists of three main modules: test case generation, weakness identification and in-context learning (ICL) demonstration. In particular, the test case generation module utilizes atomic skills defined on the two formal logic systems and an inference synthesis approach to generate questions as test cases. Then, the generated cases are fed into the LLMs to reveal weaknesses and provide insights into the LLMs by the weakness identification process. Finally, LogicAsker utilizes these insights to construct ICL demonstrations to improve the reasoning abilities of the LLMs.

3.1 Reasoning Skills

Atomic skills. As described in Section 2.1, propositional and predicate logic are two fundamental systems that formalize the reasoning process. The inference rules and equivalence laws in these two systems are atomic and can cover all correct reasoning scenarios; therefore, we define these 30 rules as the set of atomic skills an LLM should possess to perform formal reasoning.

Extended skills. Predicate logic extends propositional logic to deal with more complex statements that involve variables, quantifiers, and predicates.

Table 1: Comparison with previous works.

	Fully Automatic	Atomic Skills	Formal Rules	Include Fallacies	Identify Weakness	Improve LLMs	LLMs* Tested	Example Testbed
CLUTRR (Sinha et al., 2019)	×	×	×	×	✓	×	-	BERT
LogiQA (Liu et al., 2020)	×	×	×	×	×	×	-	BERT
RECLOR (Yu et al., 2020)	×	×	×	×	✓	×	2	GPT2
Soft Reasoner (Clark et al., 2020)	✓	×	1	×	✓	×	-	RoBERTa
LogicNLI (Tian et al., 2021)	×	×	7	×	✓	×	-	BERT
FOLIO (Han et al., 2022)	×	×	×	×	×	×	4	GPT3
LogicInference (Ontañón et al., 2022)	✓	×	19	×	×	×	-	T5
ProntoQA-OOD (Saparov et al., 2023)	✓	×	6	×	✓	×	4	GPT3.5
LogicAsker (Ours)	✓	✓	30	✓	✓	✓	6	GPT4

* We consider language models with more than 1 billion parameters as LLMs.

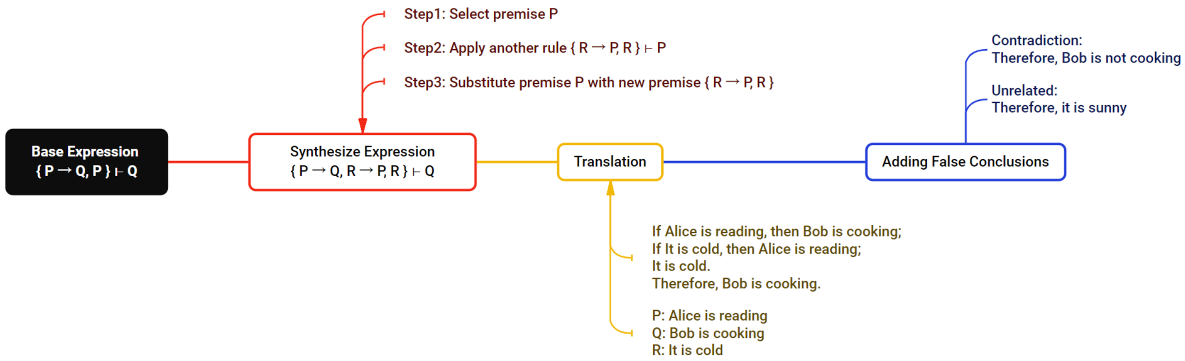

Figure 2: Test case generation procedure.

In this regard, besides the unique equivalence and inference laws in predicate logic, we add quantifiers and variables to every rule in propositional logic to form the predicate version of the laws. Using this approach, we expand the set of 30 atomic skills into a set of 208 extended skills. In Appendix B, we provide some concrete examples of these extended rules.

3.2 Test Case Generation

To generate logical questions, LogicAsker first adopts a rule-based method to generate logical expressions systematically based on reasoning skills and then translates the logical expressions into natural language. Figure 2 provides an overview of the procedure.

Logic expression generation. To better control the process of logic expression generation, we first define the length of an inference problem by the number of syllogisms it involves. We use the inference rules described in Section 2.1 to generate inference expressions with length one. When a longer inference (> 1) is specified, we start with a base expression $E_0 := P_1 \wedge P_2 \rightarrow C_1$ with length one and expand the inference chain. Specifically, we substitute the premises (either or both) of the first inference with the conclusion of some other syllogism and append the premises of those syllogisms into the list of all premises. For example, we can find another syllogism $E_1 := P_3 \wedge P_4 \rightarrow P_2$ with P_2 as the conclusion and then obtain a new expression $E_{new} := P_1 \wedge P_3 \wedge P_4 \rightarrow C_1$ with the inference length of two. We can obtain inference expressions of any length by recursively expanding the inference chain as above. During the generation process, one can specify the desired rules and length to allow complete control over expected test cases.

In addition to the correct inference expression created above, we generate three kinds of false inference expressions: contradiction, unrelated, and fallacy. A contradiction is generated by negating the conclusion of a correct inference expression and an unrelated is generated by replacing the conclusion of a valid inference expression with an irrelevant statement. For example, for $E_0 := P_1 \wedge P_2 \rightarrow C_1$, a contradiction is $E_c := P_1 \wedge P_2 \rightarrow \neg C_1$, an unrelated can be $E_u := P_1 \wedge P_2 \rightarrow U_1$. We create a fallacy by directly using the fallacy rules listed in Section 2.1 for an inference length of one. For a fallacy with a more extended length, we select a fallacy rule as the base expression and expand the inference chain using correct rules, ensuring the expression's incorrectness.

Natural language translation. Partially inspired by (Ontañón et al., 2022), translating a clause into natural language involves a series of patterns that depend on the structure of the clause. Simple propositions are transformed into one of the template patterns, such as “subject verb-action”, “subject predicate”, or “impersonal-action” with a predefined set of subjects, verbs, predicates, and impersonal actions that can be chosen randomly without repetition. For predicate clauses that involve constants or variables, we employ templates “subject verb-action”, “subject predicate” to translate them. Furthermore, each clause can be rendered in various modes, such as the present, past, or negated forms. Additionally, connectives like "or," "and," "implies," and "if and only if" also adhere to their designated patterns. For quantified clauses, we adopt patterns like "for all x, X", "there is at least one x for which X", and "some Xs are Y,". To facilitate the generation process, we curate extensive lists of potential subjects, including common names in English, and compile plausible predicates, actions, and impersonal actions. We provide a detailed illustration of the translation process in Appendix C.

3.3 Weakness Identification

To measure the reasoning abilities of the LLMs, we calculate the accuracy of LLMs' answers Acc $= \frac{N_{\text{correct}}}{N_{\text{total}}}$. Where N_{total} denotes the total number of responses, and N_{correct} denotes the number of responses that are correct. In particular, since all generated queries are formulated as yes-or-no questions, LogicAsker adopts an automatic approach that searches for pre-defined keywords (e.g., "yes" and "no") in sentences to identify correct answers.

To reveal the weaknesses of LLMs, we generate n test cases for each leaf node in the rule tree depicted in Figure 1. Then, we calculated the response accuracy of an LLM of each leaf node. Based on the result, we can identify the weaknesses of LLMs by listing the leaf nodes that receive the lowest accuracy. In addition, by grouping the accuracy by different attributes in the rule tree, we can gain insights into the strengths and weaknesses of LLMs on these attributes (e.g., performance on predicate logic vs. propositional logic).

3.4 Improving LLMs

In-context learning (ICL) is a paradigm that enables LLMs to learn tasks with examples in the form of demonstrations (Brown et al., 2020). It leverages task instructions and a few demonstration examples to convey the task semantics, which are then combined with query questions to create inputs for the language model to make predictions. ICL has demonstrated impressive performance in various natural language processing and code intelligence. However, the performance of ICL is known to rely on high-quality demonstrations (Gao et al., 2023b) strongly. To fully unleash the potential of ICL, LogicAsker utilizes the weak skills of each LLM to construct both correct and incorrect examples with expected answers and explanations as demonstrations to facilitate the reasoning of LLMs. The generation process follows a similar approach to the test case generation described in § 3.2, with the difference being that we append a brief explanation and the correct answer at the end of each case. We show an instance of the demonstration example in Appendix D.

Fine-tuning is another widely used technique to enhance model performance on specific tasks (Moslem et al., 2023; Wei et al., 2021). This process involves taking a pre-trained model and further training it on a smaller, task-specific dataset. The rationale behind fine-tuning is to leverage the learned features and knowledge of the pre-trained model, adapting it to particular nuances and characteristics of a targeted domain or task. In this paper, we directly utilize the data generated by LogicAsker to fine-tune LLMs to improve their reasoning ability.

Table 2: Conversational LLMs used in the evaluation.

Name	Rank
GPT-4o (OpenAI, 2024c)	1
GPT-4 (OpenAI, 2024b)	4
ChatGPT (OpenAI, 2024a)	37
Gemini-1.5 (Google, 2024)	10
Llama3-70b (Meta Platforms, 2024)	12
Mixtral-8x7b (Mistral AI, 2024)	42

4 Experiments

4.1 Experimental Setup

We apply LogicAsker to test six state-of-the-art (SOTA) LLMs, including four close-source and two open-source models. Table 2 lists brief information on these systems. All of them are ranked within the top 50 in the LMSYS Chatbot Arena Leaderboard [2] according to the assessment results in June 2024. We leave details of how we access

[2] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

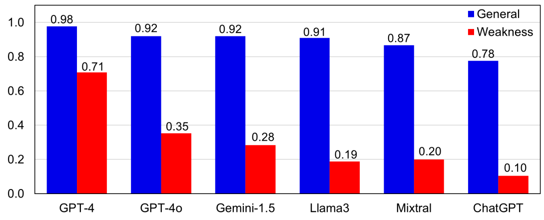

Figure 3: Overall accuracy.

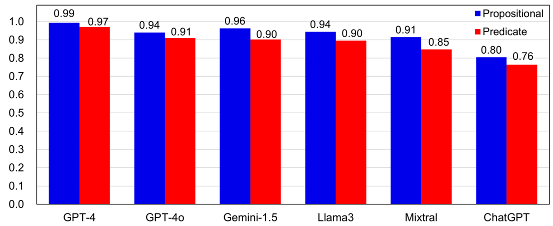

Figure 4: Propositional and predicate logic accuracy.

the model, the parameters used, and the prompt we used in Appendix E.

We conduct two iterations of experiments for a comprehensive assessment. In the first iteration, we follow the setting in § 3.3 and set $n = 25$, resulting in 5,200 cases. Statistics of the sampled data are described in Appendix **??**. The second iteration is based on the first one, which focuses on the identified weaknesses of each LLM, i.e., the ten leaf nodes in Figure 1 with the lowest accuracy. We generated 25 additional test cases for each weakness. These 250 test cases comprise our "weakness dataset, " which will be utilized for further evaluation in 4.5.

4.2 Effectiveness of LogicAsker

We demonstrate the effectiveness of LogicAsker through the overall performance of LLMs on the test cases. The overall performance of LLMs in the first and second iteration is shown in Figure 3. The result reveals that our framework can effectively expose logical failures in the first iteration, with LLM's accuracy ranging from 78%-98%. When focusing on the weak skills of LLMs in the second iteration, we further reduce the accuracy to 10%-71% for the LLMs. What's surprising is that most of these LLMs achieved response accuracy even lower than random guesses (i.e., 50% here) when confronted with logical questions involving specific logical rules. This contradicts their remarkable performance in various LLM benchmarks, for example, achieving top 50 ranks on the LLM Arena Leaderboard. It suggests that existing benchmark datasets are not comprehensive enough to assess the generalization ability of LLMs in reasoning.

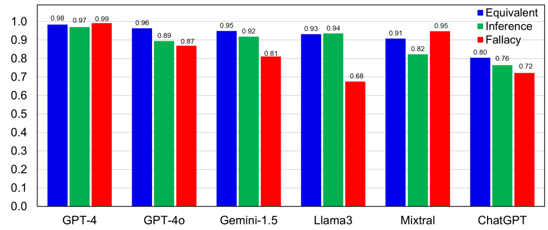

Figure 5: Accuracy of different rule categories.

4.3 Insights into Reasoning Abilities

We conducted a comprehensive analysis to gain insights from the failures exposed by LogicAsker, obtaining three key observations from the evaluation:

Most LLMs are better at easier logical skills. We compared the performance of LLMs on propositional logic and predicate logic, the former of which is simper in form while the latter involves more complex quantifier manipulations. Figure 4 illustrates the difference between the accuracy and response scores obtained for the two logic systems. A positive value indicates a higher score in propositional logic, while a negative value indicates higher scores in predicate logic. Notably, we observed that most LLMs are better at propositional logic, implying their limited ability in complex reasoning scenarios.

Most LLMs are weak in recognizing logical fallacies. Figure 5 presents the accuracy of LLMs under different skill categories. Interestingly, we discovered that among three types of skills, recognizing fallacies has the lowest accuracy for most LLMs, with GPT-4 and Mixtral-8x7B being the exceptions. It suggests that current LLMs are overconfident even in fallacies, which may be learned from the mistakes in pretraining data.

Case study: GPT-4 did not learn all logic rules well. To provide a direct impression of what skills LLMs cannot perform well, we list three atomic rules in which GPT-4 has the lowest accuracy in Table 3. While GPT-4 has an average accuracy of 98% over all skills, it only achieves 60% - 68% accuracy on these skills, indicating that it cannot perform these atomic skills smoothly.

These insights provide a valuable understanding of the strengths and weaknesses of each LLM when handling logical questions, allowing us to uncover specific areas that require improvement and potential avenues for enhancing overall performance. We provide a full breakdown list of the LLMs' performance on various skills in Appendix G.

Table 3: Weakness of GPT-4

Rule	Type	Example	Accuracy
Existential resolution	Incorrect	For all v, v will not play squash or v will go running. There is at least one v for which v will play squash or v will play tennis. Therefore, there is at least one v for which v will go running.	0.60
Universal resolution	Correct	For all x, x will climb a mountain or x is a police officer. For all x, x will not climb a mountain or x is rich. Therefore, for all x, x is a police officer or x is rich.	0.60
Law of quantifier movement	Correct	For all x, if Joseph sleeps, then x is a janitor. Therefore, if Joseph sleeps, then for all x, x is a janitor.	0.68

Table 4: Human Evaluation Results on the Quality of test cases.

Invalid Cases	a	b	c	Total
Count	10	8	0	18
Percentage	1.92%	1.54%	0.00%	3.46%

Table 5: Performance of ICL demonstrations by LogicAsker (%).

Model	Zero-Shot	CoT	ICL	ICL(Weak)
GPT-4	97.75	96.60	97.98	99.48
GPT-4o	91.92	92.94	95.77	97.23
Gemini	92.06	93.62	96.13	96.67
Llama-3	91.02	94.54	94.83	93.35
Mixtral	86.77	86.23	76.40	82.02
ChatGPT	77.62	78.19	82.90	81.04
Average	89.52	90.35	90.67	**91.63**

4.4 The Quality of Test Cases

Since the test cases are automatically generated, we conduct a human evaluation to measure the quality of the generated test cases by LogicAsker. To achieve this, we randomly sampled 10% (520) of the test cases generated during the first iteration of the experiment in 4.2 and conduct manual inspection. Two annotators with bachelor's degrees were recruited to answer the questions manually. Each test case was annotated as either valid or invalid based on the following three questions: **a)** Is the question grammatically correct? **b)** Is the question understandable and has only one interpretation? **c)** Can the target answer be derived from the question? A test case is considered valid only when both annotator's answer to the above questions are positive. The results of the annotation are presented in Table 4. This result is statistically sufficient to prove that the probability of LogicAsker generating understandable and solvable logical questions is larger than or equal to 0.94 (with p-value 0.05), indicating that **the test cases generated by LogicAsker are highly reliable and valid**.

4.5 LogicAsker to Improve Reasoning

In this section, we explore the potential of LogicAsker in further improving the reasoning ability of LLMs through in-context learning (ICL) and fine-tuning.

We first employ LogicAsker to generate ICL demonstrations tailored to address the weaknesses dataset uncovered in the experiments of 4.2. For each inference problem, we generated ICL demonstrations that provide both the expected answer and an explanation as described in § 3. We evaluate the effectiveness of the ICL demonstrations generated by LogicAsker by comparing the following prompting strategies: a) Zero-Shot: We provide only task instructions without any ICL demonstrations. b) Zero-Shot Chain-of-Thouhgt (CoT): We use the instruction "Please think step-by-step" (Kojima et al., 2022) to elicit the zero-shot reasoning ability of the LLMs. c) Random ICL Demonstrations: In addition to the task instruction, we also include four ICL demonstrations selected randomly from the available rules with balanced answer labels, i.e., two correct and two incorrect. d) Weakness ICL Demonstration: Instead of random demonstrations, we include four ICL demonstrations using the weakness rules identified in 4.2 with balanced answer labels.

We perform testing with the 5.2k sampled data on all models and list the result in Table 5. In general, the weakness ICL demonstrations are more effective than those random ICL demonstrations, and both ICL methods bring more performance gain than CoT, **indicating that the test cases generated by LogicAsker can improve reasoning**.

To further demonstrate the effectiveness of LogicAsker, we fine-tune ChatGPT on 5.2k separately generated data on all skills and 2.8k separately generated data on weaknesses of ChatGPT, respectively. We test the two fine-tuned model on both LogicAsker and another dataset, LogiQA, a challenging dataset for machine reading comprehension with logical reasoning (Liu et al., 2020). We use the "test" split of LogiQA which contains 651 test data. The results are presented in Table 6. We can observe that models fine-tuned on LogicAsker can effectively enhance the models' reasoning ability on both datasets, suggesting the generalizability of LogicAsker. **These findings demonstrate the effectiveness of LogicAsker in improving the reasoning ability of LLMs**.

Table 6: ChatGPT performance on LogiQA and LogicAsker after fine-tuning (%).

	Vanilla	FT (All)	FT (Weak)
LogicAsker	77.62	**99.50**	97.83
LogiQA	40.55	41.01	**41.78**

5 Related Work

Significant advancements in NLP reasoning have been achieved through methods such as Chain-of-Thoughts (CoT) prompting (Wei et al., 2022b), which enables models to generate reasoning steps with minimal training. Enhancing this, the Program of Thoughts (PoT) prompting (Chen et al., 2022) leverages external interpreters like Python to handle complex computations, demonstrating substantial improvements over CoT in tackling intricate mathematical problems.

Recent studies have focused on evaluating the reasoning capabilities of Large Language Models (LLMs) by measuring their performance across various reasoning tasks. These include arithmetic (Cobbe et al., 2021; Hendrycks et al., 2021; Amini et al., 2019; Patel et al., 2021; Miao et al., 2020; Ling et al., 2017; Roy and Roth, 2016), commonsense (Talmor et al., 2019; Geva et al., 2021; Clark et al., 2018), symbolic (Wei et al., 2022b), and table reasoning (Nan et al., 2021), as well as understanding words, dates, and causal relationships (Aarohi Srivastava, 2022), and generalization (Lake and Baroni, 2017; Anil et al., 2022). Despite these efforts, it remains uncertain whether LLMs truly reason or rely on simple heuristics, since most assessments focus only on accuracy and do not thoroughly evaluate the reasoning processes.

Efforts to develop metrics for more formal reasoning analysis in LLMs include creating datasets with first-order logic problems (Han et al., 2022), generating test cases using a single predicate inference rule (Saparov and He, 2022), and employing propositional logic with randomized methods (Ontañón et al., 2022). These methods, however, often lack generalizability or focus on limited deduction rules. (Saparov et al., 2023) introduced a comprehensive approach by using all deduction rules in propositional logic to assess LLMs' deductive reasoning across complex proofs. Our research expands further, incorporating all rules and equivalent laws in both propositional and predicate logic, aiming to enhance understanding of each rule's impact on LLM performance and using these insights for improvement.

6 Conclusion

In this paper, we present LogicAsker, an automated tool designed to comprehensively evaluate and improve the formal reasoning abilities of LLMs under a set of atomic skills.

Our research demonstrated the efficacy of LogicAsker in identifying logical reasoning failures in a diverse set of widely deployed LLMs, we achieved a substantial success rate in revealing reasoning flaws in these models, ranging from 25% to 94%. Additionally, we utilized the test cases from LogicAsker to design in-context learning demonstrations, which effectively enhance the logical reasoning capabilities of LLMs, e.g., improving from 75% to 85% for GPT-4.

By providing insights into the strengths and weaknesses of LLMs in reasoning, we are able to improve the reliability and trustworthiness of these models. The release of all the code and data aims to facilitate replication and encourage further research in this crucial area.

Limitations

This paper identifies two primary limitations that highlight areas for future research:

- Although our ICL (In-Context Learning) method significantly enhances the logical reasoning capabilities of large language models (LLMs), there remains a performance gap compared to human-level reasoning. Further refinements and innovations in model training and architecture may be necessary to bridge this gap.

- Our method is currently applicable only to LLMs that possess robust in-context learning capabilities. LLMs lacking this feature may not benefit from our approach. Future studies could explore fine-tuning methods to extend the applicability of our improvements across a broader spectrum of LLMs, potentially enhancing models with weaker or no inherent in-context learning abilities.

References

et al. Aarohi Srivastava. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *ArXiv*, abs/1905.13319.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Venkatesh Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. *ArXiv*, abs/2207.04901.

Hugo Bronkhorst, Gerrit Roorda, Cor J. M. Suhre, and Martin J. Goedhart. 2020. Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18:1673–1694.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *ArXiv*, abs/2211.12588.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *International Joint Conference on Artificial Intelligence*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2022. Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*.

Shuzheng Gao, Xinjie Wen, Cuiyun Gao, Wenxuan Wang, and Michael R. Lyu. 2023a. Constructing effective in-context demonstration for code intelligence tasks: An empirical study. *ArXiv*, abs/2304.07575.

Shuzheng Gao, Xinjie Wen, Cuiyun Gao, Wenxuan Wang, and Michael R. Lyu. 2023b. What makes good in-context demonstrations for code intelligence tasks with llms? *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 761–773.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Google. 2024. Gemini. `https://gemini.google.com/`. Accessed: 2024-06-16.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq R. Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir R. Radev. 2022. Folio: Natural language reasoning with first-order logic. *ArXiv*, abs/2209.00840.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *ArXiv*, abs/2212.10403.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *ArXiv*, abs/2310.01798.

Hugging Face. 2024. Chatbot arena leaderboard. `https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard`. Accessed: 2024-06-16.

Patrick J. Hurley and Lori Watson. 2020. A concise introduction to logic, 13/e.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *ArXiv*, abs/2301.08745.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NeurIPS*.

Brenden M. Lake and Marco Baroni. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Annual Meeting of the Association for Computational Linguistics*.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yuexin Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4.

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

Meta Platforms. 2024. Llama-3. `https://llama.meta.com/llama3/`. Accessed: 2024-06-16.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing english math word problem solvers. *ArXiv*, abs/2106.15772.

Mistral AI. 2024. Mixtral of experts. `https://mistral.ai/news/mixtral-of-experts/`. Accessed: 2024-06-16.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Fine-tuning large language models for adaptive machine translation. *ArXiv*, abs/2312.12740.

Linyong Nan, Chia-Hsuan Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Benjamin Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2021. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Santiago Ontañón, Joshua Ainslie, Vaclav Cvicek, and Zachary Kenneth Fisher. 2022. Logicinference: A new dataset for teaching logical inference to seq2seq models. *ArXiv*, abs/2203.15099.

OpenAI. 2024a. Chatgpt. `https://openai.com/chatgpt/`. Accessed: 2024-06-16.

OpenAI. 2024b. Gpt-4. `https://openai.com/index/gpt-4/`. Accessed: 2024-06-16.

OpenAI. 2024c. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-06-16.

Barbara H. Partee, Alice ter Meulen, and Robert E. Wall. 1990. Mathematical methods in linguistics.

Arkil Patel, S. Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *North American Chapter of the Association for Computational Linguistics*.

Marco Tulio Ribeiro, Tongshuang Sherry Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *ACL*.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *ArXiv*, abs/1608.01413.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *ArXiv*, abs/2210.01240.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using ood examples. *ArXiv*, abs/2305.15269.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. In *Conference on Empirical Methods in Natural Language Processing*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *ArXiv*, abs/1811.00937.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Maarten Bosma, Vincent 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. *ArXiv*, abs/2109.01652.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *ArXiv*, abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *NeurIPS*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *ArXiv*, abs/2002.04326.

A Logical Rules and Fallacies

We list all the logic equivalence rules in Table 7-8, logic inference rules in Table 9, and common logical fallacies in Table 10.

B Extended Rules

B.1 Equivalent Extension

The equivalent rule extension is based on the following fact:

$$\{A \Leftrightarrow B, \forall x(A)\} \vdash \{\forall x(B)\}$$

(i.e., if A and B are equivalent, and for all x, A is true, then for all x, B is also true), and

$$\{A \Leftrightarrow B, \exists x(A)\} \vdash \{\exists x(B)\}$$

(i.e., if A and B are equivalent, and there exist x such that A is true, then there exist x such that B is true). For example, the predicate version of the DeMorgan's law

$$\neg(P \wedge Q) \Leftrightarrow \neg P \vee \neg Q$$

will become

$$\forall x(\neg(P(x) \wedge Q(x))) \Leftrightarrow \forall x(\neg P(x) \vee \neg Q(x)),$$

and

$$\exists x(\neg(P(x) \wedge Q(x))) \Leftrightarrow \exists x(\neg P(x) \vee \neg Q(x)).$$

In this example, the goal is to extend the propositional equivalence law to its predicate version by adding quantifiers. To achieve this goal, we first note that DeMorgan's law states that "P and Q cannot both be true" (e.g., Alice is happy and Bob is happy cannot both be true) is equivalent to "either not P or not Q" (e.g., either Alice is not happy or Bob is not happy). Since the two expressions are equivalent, we can add the same quantifier to both sides and the equivalence will still hold. Therefore, by adding a "for all" quantifier to both sides, we obtain "for all x, P(x) and Q(x) cannot both be true" (for all persons in the room, the person likes Charley and the person likes David cannot both be true) is equivalent to "for all x, either not P(x) or not Q(x)" (e.g., for all person in the room, either the person doesn't like Charley or the person doesn't like David). Before the extension, the law can only be applied to simple propositions (e.g., P = "Alice is happy", Q = "Bob is happy"), but after extension, the law can be applied to predicates with variables and quantifiers (e.g., P(x) = "x likes Charley", Q(x) = "x likes David") The same also applies to the "exist" quantifier.

B.2 Inference Extension

The inference rule extension is based on the following fact:

$$\{A \wedge B \rightarrow C\} \vdash \{\forall x, (A) \wedge \forall x, (B) \rightarrow \forall x, (C)\},$$

(i.e., if A and B imply C, then for all x, A is true and for all x, B is true implies for all x, C is true)

$$\{A \wedge B \rightarrow C\} \vdash \{\exists x, (A) \wedge \forall x, (B) \rightarrow \exists x, (C)\}.$$

(i.e., if A and B imply C, there exists x such that A is true and for all x, B is true implies there exists x such that C is true). Since all propositional inference rules are of the form $P \wedge Q \rightarrow C$, we can transform them into their predicate form $\forall x, P(x) \wedge \forall x, Q(x) \rightarrow \forall x, C(x)$ and $\exists x, P(x) \wedge \forall x, Q(x) \rightarrow \exists x, C(x)$ following similar procedure in the previous section.

C Natural Language Translation

C.1 Algorithm

Given an input: a logic clause of the form `[operator, Clause`$_A$`, Clause`$_B$`]`, where the clauses are also of the form `[operator, Clause`$_A$`, Clause`$_B$`]`, the algorithm will do the following:

1. **Single Proposition Clause:** If the clause is just a single proposition, the algorithm finds this proposition's natural language form and returns it. The natural language form is obtained by combining vocabularies according to certain templates (e.g., subject + action).

2. **Negation:** If the clause starts with a "$\neg$" operator, the algorithm then translates the rest of the clause based on a negation template, making sure to negate the statement.

3. **Quantifiers:** For clauses that start with "$\forall$" (meaning for all items) or "$\exists$" (meaning there is at least one item), it translates these into natural language, adjusting the phrasing based on whether we're asserting something positively or negating it.

4. **Logical Connectives:** If the clause combines propositions using logical operators like "$\wedge$", "$\vee$", "$\rightarrow$" (implies), or "$\leftrightarrow$" (if and only if), the function translates these into natural language phrases that express the relationship between the propositions.

Table 7: Propositional logic equivalence laws.

Law	Logical Equivalence	Example
Idempotent laws	$P \wedge P \Leftrightarrow P$ $P \vee P \Leftrightarrow P$	I am a teacher and I am a teacher ⇔ I am a teacher. It's raining or it's raining ⇔ it's raining.
Commutative laws	$P \wedge Q \Leftrightarrow Q \wedge P$ $P \vee Q \Leftrightarrow Q \vee P$	It is cold and it is winter ⇔ It is winter and it is cold. You can go to the party or you can study ⇔ You can study or you can go to the party.
Associative laws	$(P \wedge Q) \wedge R \Leftrightarrow P \wedge (Q \wedge R)$ $(P \vee Q) \vee R \Leftrightarrow P \vee (Q \vee R)$	It is raining and it is cold, and also it is winter ⇔ It is raining, and also, it is cold and it is winter. Either I will go to the park or I will go to the library is true, or I will go to the cinema ⇔ I will go to the park or either I will go to the library or I will go to the cinema is true.
Distributive laws	$P \wedge (Q \vee R) \Leftrightarrow (P \wedge Q) \vee (P \wedge R)$ $P \vee (Q \wedge R) \Leftrightarrow (P \vee Q) \wedge (P \vee R)$	It is raining and either I have an umbrella or I have a raincoat ⇔ It is raining and I have an umbrella, or it is raining and I have a raincoat. Either I will go to the park, or it is cloudy and it is cold ⇔ Either I will go to the park or it is cloudy is true, and either I will go to the park or it is cold is true.
DeMorgan's laws	$\neg(P \wedge Q) \Leftrightarrow \neg P \vee \neg Q$ $\neg(P \vee Q) \Leftrightarrow \neg P \wedge \neg Q$	It is not true that it's both cold and raining ⇔ It's not cold or it's not raining. It's not true that I will study or play ⇔ I won't study and I won't play.
Complement laws	$\neg(\neg P) \Leftrightarrow P$ $P \wedge \neg P \Leftrightarrow False$ $P \vee \neg P \Leftrightarrow True$	It is not the case that it is not raining ⇔ It is raining. It is raining and it is not raining. It is raining or it is not raining.
Conditional laws	$P \rightarrow Q \Leftrightarrow \neg P \vee Q$	If it rains, then I'll stay at home ⇔ It doesn't rain or I stay at home.
Bidirectional laws	$(P \leftrightarrow Q) \Leftrightarrow (P \wedge Q) \vee (\neg P \wedge \neg Q)$	I'll go to the park if and only if it's sunny ⇔ Either it's sunny and I go to the park, or it's not sunny and I don't go to the park.
Identity laws	$P \wedge True \Leftrightarrow P$ $P \vee False \Leftrightarrow P$	It is raining and it is true ⇔ It is raining. I will study or it's false ⇔ I will study.

C.2 Example

Consider the expression: `[`$\forall x$`, `$\rightarrow$`, A(`x`), B(`x`)]`. Here's how the function would translate it:

1. It sees the "$\forall x$" quantifier and adds "For all x," to the sentence and continues to process the clause `[`$\rightarrow$`, A(`x`), B(`x`)]`.

2. It sees the "$\rightarrow$" operator, which means "if...then...". It connects the two operands with the operator and obtains "For all x, if A(x), then B(x)". Then, it continues to process the clauses A(x), B(x).

3. Since A(x), B(x) are single proposition clauses, the function looks up the vocabulary and synthesizes the natural language versions of the proposition. For example, A(x) = "x drinks water", B(x) = "x is a cashier".

4. It constructs the sentence: "For all x, if x drinks water, then x is a cashier".

C.3 Vocabulary

We list the vocabulary used in our experiment:

Subjects

- x, y, z, James, Mary, Robert, Patricia, John, Jennifer, Michael, Linda, William, Elisabeth, David, Barbara, Richard, Susan, Joseph, Jessica, Thomas, Sarah, Charles, Karen, Alice, Benjamin, Daniel, Emily, George, Helen, Ian, Julie.

Predicates

- a cashier, a janitor, a bartender, a server, an office clerk, a mechanic, a carpenter, an electrician, a nurse, a doctor, a police officer, a taxi driver, a soldier, a politician, a lawyer, a scientist, an astronaut, a poet, an artist, a sailor, a writer, a musician, poor, rich, happy, sad, fast, curious, excited, bored, tired, joyful, intelligent, skilled, efficient, meticulous, creative.

Actions

- make tea, makes tea, making tea, drink water, drinks water, drinking water, read a book, reads a book, reading a book, play tennis,

Table 8: Predicate logic quantifier laws.

Law	Logical Equivalence	Example
Quantifier Negation	$\neg\forall x P(x) \Leftrightarrow \exists x \neg P(x)$	It is not the case that all birds can fly ⇔ There exists a bird that cannot fly.
	$\neg\exists x P(x) \Leftrightarrow \forall x \neg P(x)$	There is no human that can live forever ⇔ All humans cannot live forever.
Quantifier Distribution	$\forall x(P(x) \wedge Q(x)) \Leftrightarrow \forall x P(x) \wedge \forall x Q(x)$	Every student is smart and diligent ⇔ Every student is smart, and every student is diligent.
	$\exists x(P(x) \vee Q(x)) \Leftrightarrow \exists x P(x) \vee \exists x Q(x)$	There is a person who is either a doctor or a lawyer ⇔ There is a person who is a doctor, or there is a person who is a lawyer.
Quantifier Commutation	$\exists x \exists y P(x, y) \Leftrightarrow \exists y \exists x P(x, y)$	There exists a child and a toy such that the child owns the toy ⇔ There exists a toy and a child such that the child owns the toy.
	$\forall x \forall y P(x, y) \Leftrightarrow \forall y \forall x P(x, y)$	For all parents and children, the parent loves the child ⇔ For all children and parents, the parent loves the child.
Quantifier Transposition	$\exists x \forall y P(x, y) \Leftrightarrow \forall y \exists x P(x, y)$	There exists a food that all people like is equivalent to For all people, there exists a food that they like.
Quantifier Movement	$\forall x(P \rightarrow Q(x)) \Leftrightarrow (P \rightarrow \forall x Q(x))$	For every child, if it is raining then they are inside ⇔ If it is raining, then every child is inside when the notion of raining doesn't depend on the specific child.
	$\exists x(P \wedge Q(x)) \Leftrightarrow (P \wedge \exists x Q(x))$	There exists a student who is tall and a good basketball player ⇔ There is a tall student and there exists a student who is a good basketball player when the notion of being tall doesn't depend on the specific student.

plays tennis, playing tennis, play squash, plays squash, playing squash, play a game, plays a game, playing a game, go running, goes running, running, work, works, working, sleep, sleeps, sleeping, cook, cooks, cooking, listen to a song, listens to a song, listening to a song, write a letter, writes a letter, writing a letter, drive a car, drives a car, driving a car, climb a mountain, climbs a mountain, climbing a mountain, take a plane, takes a plane, taking a plane, paint a picture, paints a picture, painting a picture.

Impersonal Candidates

- snowing, snows, doesn't snow, snow, raining, rains, doesn't rain, rain, sunny, is sunny, is not sunny, be sunny, cloudy, is cloudy, is not cloudy, be cloudy, windy, is windy, is not windy, be windy, cold, is cold, is not cold, be cold, late, is late, is not late, be late, overcast, is overcast, is not overcast, be overcast, foggy, is foggy, is not foggy, be foggy, humid, is humid, is not humid, be humid.

D Prompting LLMs

For all GPT models, we set the system prompt of to blank.

D.1 Zero-Shot Example

```
Consider the following premises: The claim that John is a poet and the claim that it is cloudy cannot both be true. Can we infer the following from them? Answer yes or no: Jessica is not listening to a song.
```

D.2 Zero-Shot CoT

For Zero-Shot CoT, we add `"Please think step-by-step and answer the following question."` To zero-shot queries.

D.3 ICL Example

```
Q: Consider the following premises: If it snows, then Joseph is a politician. Can we infer the following from them? Answer yes or no: It is snowing if and only if Joseph is a politician.

A: Let A be the claim that "it snows", B be the claim that "Joseph is a politician", then the premises are "if A then B, if B then A", which is equivalent to "A if and only if B" by the biconditional introduction rule. Therefore, we can infer that It is snowing if and only if
```

Table 9: Propositional and predicate logic inference rules.

Inference Rule	Logical Form	Example
Universal Instantiation	$\forall x P(x) \vdash P(c)$	All birds have wings. Hence, this crow has wings.
Existential Generalization	$P(c) \vdash \exists x P(x)$	This apple is red. Hence, there exists a red apple.
Modus Ponens	$\{P \to Q, P\} \vdash Q$	If it rains, the street gets wet. It is raining. Hence, the street is wet.
Modus Tollens	$\{P \to Q, \neg Q\} \vdash \neg P$	If I study, I will pass the test. I did not pass the test. Hence, I did not study.
Transitivity	$\{P \to Q, Q \to R\} \vdash P \to R$	If it rains, I take my umbrella. If I take my umbrella, I won't get wet. Hence, if it rains, I won't get wet.
Disjunctive Syllogism	$\{P \vee Q, \neg P\} \vdash Q$	Either it's raining or it's snowing. It's not raining. Hence, it's snowing.
	$\{P \vee Q, \neg Q\} \vdash P$	Either it's raining or it's snowing. It's not snowing. Hence, it's raining.
Addition	$\{P\} \vdash P \vee Q$	It is raining. Hence, it is raining or it is snowing.
	$\{Q\} \vdash P \vee Q$	It is snowing. Hence, it is raining or it is snowing.
Simplification	$\{P \wedge Q\} \vdash P$	It is raining and it is cold. Hence, it is raining.
	$\{P \wedge Q\} \vdash Q$	It is raining and it is cold. Hence, it is cold.
Conjunction	$\{P, Q\} \vdash P \wedge Q$	It is raining. It is cold. Hence, it is raining and it is cold.
Constructive Dilemma	$\{P \to Q, R \to S, P \vee R\} \vdash Q \vee S$	If it rains, I'll stay at home. If I work, I'll be tired. Either it will rain or I'll work. Hence, I'll either stay at home or be tired.

Table 10: Common fallacies.

Name	Logical Form	Example
Affirming the Consequent	$p \to q, q \vdash p$	If I study, I will pass the test. I passed the test. Therefore, I studied.
Denying the Antecedent	$p \to q, \neg p \vdash \neg q$	If it rains, the street gets wet. It is not raining. Therefore, the street is not wet.
Affirming a Disjunct	$p \vee q, p \vdash \neg q$	Either I will study or I will fail the test. I studied. Therefore, I will not fail the test.
Denying a Conjunct	$\neg(p \wedge q), \neg p \vdash q$	I'm not both hungry and thirsty. I'm not hungry. Therefore, I'm thirsty.
Illicit Commutativity	$p \to q \vdash q \to p$	If I am in Paris, then I am in France. Therefore, if I am in France, I am in Paris.
Undistributed Middle	$\forall x(P(x) \to Q(x)), Q(a) \vdash P(a)$	All dogs are animals. My cat is an animal. Therefore, my cat is a dog.

```
Joseph is a politician. The answer is
==yes==.

Q: Consider the following premises:
It is late and it is windy. Can we infer
the following from them? Answer yes or
no: It is windy.

A: Let A be the claim that "it is
late", B be the claim that "it is windy",
then the premises are "A and B". By
the simplification rule, we can infer B.
Therefore, we can infer that it is windy.
The answer is ==yes==.

Q: Consider the following premises:
Jessica is not running. Jessica is
running or it is raining. Can we infer
the following from them? Answer yes or
no: Sarah is not happy.

A: Let A be the claim that "Jessica
is running", B be the claim that "it
is raining", C be the claim that "Sarah
is happy", then the premises are "not
A, A or B". We cannot infer C from the
premises. The answer is ==no==.

Q: Consider the following premises:
It is not raining. It is raining or it
is late. Can we infer the following from
them? Answer yes or no: It is not late.

A: Let A be the claim that "it is
raining", B be the claim that "it is
late", then the premises are "not A, A
or B". We can infer "B", which is "it is
late". Therefore, we cannot infer "it is
```

```
not late". The answer is ==no==.

Q: Consider the following premises:
For all x, x will write a letter, and x
will climb a mountain and x is a musician.
Can we infer the following from them?
Answer yes or no: For all x, x will write
a letter and x will climb a mountain, and
x is a musician.
```

E Accessing LLMs

For commercial models, the specific models we accessed are `gpt-4o-2024-05-13` for GPT-4o, `gpt-4-turbo-2024-04-09` for GPT-4, `gpt-3.5-turbo-0125` for ChatGPT, `gemini-1.5-flash-latest` (accessed June, 2024) for Gemini-1.5. All accesses are made via their official APIs [3]. For the open-source models, we use their respective Hugging Face [4] repository, i.e., `meta-llama/Meta-Llama-3-70B` for Llama-3 and `mistralai/Mixtral-8x7B-v0.1` for Mixtral-8x7B.

For model parameters, we set the *temperature* to 0.0 and *max_tokens* to 500 for all models. We keep other parameters to the models' respective default values.

F Sampled Data Statistics

We analyzed various statistics of the sampled dataset, classified into different categories as presented in Tables 11, 12, and 13. Table 11 summarizes the distribution of logic types, Table 12 categorizes the rule types used in our analysis, and Table 13 details the types of problems.

Table 11: Distribution of Logic Types

Logic Type	Count
Predicate	146
Propositional	62

G Complete Break-Down Result for All LLMs.

In this section, we list the complete results of all LLMs on the set of 208 atomic rules in Table 14 through Table 19. The results are sorted in ascending order according to the zero-shot performance of the models.

[3]https://platform.openai.com/docs/ and https://ai.google.dev/gemini-api

[4]https://huggingface.co/

Table 12: Rule Categories

Rule Category	Count
Inference	108
Equivalent	81
Fallacy	19

Table 13: Problem Types

Problem Type	Count
Inference	82
Unrelated	63
Contradiction	63

Table 14: Break-down of the accuracy of GPT-4 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Universal generalization	Inference	0.52	0.52	0.52	0.52
Predicate	Inference	Existential resolution	Unrelated	0.60	0.76	1.00	1.00
Predicate	Inference	Universal resolution	Inference	0.60	0.36	0.08	1.00
Predicate	Equivalent	Law of quantifier movement	Inference	0.68	0.60	0.60	0.72
Predicate	Inference	Existential resolution	Inference	0.72	0.52	0.96	0.96
Predicate	Inference	Existential biconditional introduction	Inference	0.76	0.44	0.96	1.00
Predicate	Inference	Existential biconditional introduction	Contradiction	0.80	0.80	1.00	1.00
Predicate	Equivalent	Existential conditional laws	Inference	0.80	0.80	1.00	1.00
Predicate	Inference	Existential transitivity	Inference	0.80	0.76	0.96	1.00
Predicate	Equivalent	Universal distributive laws	Inference	0.80	0.36	0.20	0.96
Propositional	Fallacy	Denying a conjunct	Inference	0.88	0.92	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Inference	0.88	0.88	0.96	0.92
Predicate	Equivalent	Law of quantifier movement	Contradiction	0.92	1.00	1.00	0.96
Predicate	Inference	Existential biconditional elimination	Inference	0.92	0.92	1.00	1.00
Predicate	Inference	Existential biconditional introduction	Unrelated	0.92	0.96	1.00	1.00
Predicate	Equivalent	Existential biconditional laws	Contradiction	0.92	0.92	1.00	1.00
Predicate	Inference	Existential conjunction	Unrelated	0.92	0.92	0.92	0.96
Predicate	Inference	Existential modus tollens	Inference	0.92	0.88	1.00	1.00
Propositional	Inference	Resolution	Inference	0.92	0.64	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Inference	0.92	0.92	0.44	1.00
Predicate	Inference	Universal generalization	Contradiction	0.92	0.88	1.00	1.00
Propositional	Equivalent	De morgan's laws	Contradiction	0.96	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Inference	0.96	0.96	1.00	1.00
Propositional	Equivalent	Associative laws	Unrelated	0.96	0.96	1.00	1.00
Propositional	Equivalent	Conditional laws	Contradiction	0.96	0.96	1.00	1.00
Predicate	Equivalent	Existential associative laws	Unrelated	0.96	0.96	1.00	1.00
Predicate	Equivalent	Existential biconditional laws	Inference	0.96	0.96	1.00	1.00
Predicate	Fallacy	Existential denying a conjunct	Inference	0.96	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Inference	0.96	0.92	1.00	1.00
Predicate	Inference	Existential modus ponens	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Existential transitivity	Contradiction	0.96	0.92	1.00	1.00
Propositional	Equivalent	Idempotent laws	Inference	0.96	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Inference	0.96	1.00	1.00	1.00
Propositional	Inference	Simplification	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Inference	0.96	0.96	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Inference	0.96	0.88	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Contradiction	0.96	0.92	1.00	1.00
Predicate	Inference	Universal conjunction	Unrelated	0.96	0.96	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Unrelated	0.96	0.96	0.92	1.00
Predicate	Inference	Universal disjunctive syllogism	Unrelated	0.96	0.96	1.00	1.00
Predicate	Inference	Universal resolution	Unrelated	0.96	1.00	1.00	1.00
Propositional	Equivalent	De morgan's laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	De morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier movement	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Affirming a disjunct	Inference	1.00	1.00	1.00	1.00
Propositional	Fallacy	Affirming the consequent	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Associative laws	Contradiction	1.00	0.92	0.96	1.00
Propositional	Equivalent	Associative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Biconditional laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Biconditional laws	Inference	1.00	1.00	0.96	1.00
Propositional	Equivalent	Biconditional laws	Unrelated	1.00	0.96	1.00	1.00
Propositional	Equivalent	Commutative laws	Contradiction	1.00	0.96	1.00	1.00
Propositional	Equivalent	Commutative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Conditional laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Conditional laws	Unrelated	1.00	0.96	1.00	1.00
Propositional	Inference	Conjunction	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Denying the antecedent	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Inference	1.00	1.00	1.00	1.00

Continued on next page

Table 14: Break-down of the accuracy of GPT-4 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Propositional	Inference	Disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Distributive laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Distributive laws	Inference	1.00	0.96	1.00	1.00
Propositional	Equivalent	Distributive laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Existential affirming a disjunct	Inference	1.00	1.00	1.00	1.00
Predicate	Fallacy	Existential affirming the consequent	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Inference	1.00	1.00	1.00	0.96
Predicate	Inference	Existential biconditional elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential biconditional elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential biconditional laws	Unrelated	1.00	0.96	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Unrelated	1.00	0.84	1.00	1.00
Predicate	Equivalent	Existential complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential conditional laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Inference	1.00	0.96	1.00	1.00
Predicate	Fallacy	Existential denying the antecedent	Inference	1.00	0.96	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Unrelated	1.00	0.96	1.00	1.00
Predicate	Fallacy	Existential fallacy	Inference	1.00	1.00	0.96	1.00
Predicate	Inference	Existential generalization	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Existential illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Inference	1.00	0.96	1.00	1.00
Predicate	Inference	Existential modus tollens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus tollens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential resolution	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Unrelated	1.00	0.96	1.00	1.00
Predicate	Inference	Existential transitivity	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Fallacy	Illicit major	Inference	1.00	1.00	1.00	1.00
Predicate	Fallacy	Illicit minor	Inference	1.00	0.96	1.00	0.96
Propositional	Inference	Modus ponens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Resolution	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Resolution	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Simplification	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Simplification	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Transitivity	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Transitivity	Inference	1.00	1.00	0.96	1.00
Propositional	Inference	Transitivity	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Undistributed middle	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Unrelated	1.00	1.00	1.00	1.00

Continued on next page

Table 14: Break-down of the accuracy of GPT-4 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Fallacy	Universal affirming a disjunct	Inference	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal affirming the consequent	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Inference	1.00	0.92	1.00	1.00
Predicate	Equivalent	Universal associative laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Inference	1.00	0.80	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Unrelated	1.00	0.96	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Inference	1.00	1.00	0.92	1.00
Predicate	Fallacy	Universal denying a conjunct	Inference	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal denying the antecedent	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Contradiction	1.00	1.00	0.72	1.00
Predicate	Inference	Universal disjunctive syllogism	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunctive syllogism	Inference	1.00	0.88	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Inference	1.00	1.00	0.88	1.00
Predicate	Inference	Universal modus ponens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus tollens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus tollens	Inference	1.00	1.00	0.92	1.00
Predicate	Inference	Universal modus tollens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal resolution	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal simplification	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal simplification	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Unrelated	1.00	1.00	1.00	1.00

Table 15: Break-down of the accuracy of GPT-4o on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Propositional	Inference	Resolution	Inference	0.04	0.64	1.00	1.00
Predicate	Inference	Universal resolution	Inference	0.08	0.00	0.16	0.88
Predicate	Inference	Existential biconditional introduction	Unrelated	0.20	0.96	1.00	0.56
Propositional	Inference	Biconditional introduction	Unrelated	0.40	1.00	1.00	1.00
Propositional	Fallacy	Denying a conjunct	Inference	0.40	1.00	1.00	1.00
Predicate	Fallacy	Existential denying the antecedent	Inference	0.40	1.00	0.96	0.88
Predicate	Inference	Universal biconditional introduction	Unrelated	0.40	1.00	1.00	1.00
Predicate	Inference	Existential resolution	Inference	0.52	0.00	0.00	0.56
Predicate	Inference	Universal generalization	Inference	0.52	0.52	0.52	0.52
Predicate	Inference	Universal simplification	Inference	0.56	0.68	0.96	0.96
Predicate	Inference	Universal disjunction elimination	Inference	0.60	0.32	0.44	0.92
Predicate	Equivalent	Law of quantifier distribution	Inference	0.64	0.56	0.60	0.80
Predicate	Equivalent	Law of quantifier movement	Inference	0.64	0.44	0.72	0.84
Propositional	Equivalent	Conditional laws	Contradiction	0.64	0.96	0.96	0.96
Predicate	Equivalent	Existential conditional laws	Contradiction	0.64	1.00	1.00	1.00
Predicate	Inference	Existential transitivity	Inference	0.64	0.00	0.20	0.60
Predicate	Inference	Universal transitivity	Inference	0.68	0.68	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Contradiction	0.72	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Contradiction	0.72	1.00	1.00	1.00
Predicate	Fallacy	Existential fallacy	Inference	0.72	1.00	0.96	1.00
Predicate	Equivalent	Universal associative laws	Inference	0.72	0.96	0.88	1.00
Predicate	Inference	Universal disjunctive syllogism	Inference	0.72	0.88	1.00	1.00
Predicate	Inference	Existential biconditional introduction	Inference	0.76	0.04	0.16	0.28

Continued on next page

Table 15: Break-down of the accuracy of GPT-4o on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Existential resolution	Unrelated	0.76	1.00	0.96	0.64
Predicate	Fallacy	Illicit minor	Inference	0.76	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Inference	0.76	0.84	1.00	1.00
Predicate	Inference	Universal disjunctive syllogism	Contradiction	0.76	1.00	1.00	1.00
Propositional	Fallacy	Denying the antecedent	Inference	0.80	1.00	1.00	0.96
Predicate	Inference	Existential biconditional elimination	Unrelated	0.80	0.96	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Inference	0.80	0.08	0.12	0.76
Predicate	Inference	Universal biconditional elimination	Inference	0.80	0.88	0.92	1.00
Predicate	Inference	Existential biconditional elimination	Contradiction	0.84	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Unrelated	0.84	0.88	0.92	0.92
Predicate	Fallacy	Existential denying a conjunct	Inference	0.84	0.96	1.00	1.00
Predicate	Inference	Existential modus tollens	Inference	0.84	0.64	0.80	1.00
Predicate	Inference	Existential transitivity	Contradiction	0.84	1.00	1.00	0.92
Predicate	Fallacy	Universal denying the antecedent	Inference	0.84	1.00	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Inference	0.84	0.04	0.44	0.60
Predicate	Inference	Universal modus tollens	Inference	0.84	0.64	0.88	1.00
Predicate	Equivalent	Existential de morgan's laws	Inference	0.88	1.00	1.00	1.00
Predicate	Inference	Existential biconditional introduction	Contradiction	0.88	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Inference	0.88	0.96	1.00	1.00
Predicate	Inference	Existential modus tollens	Unrelated	0.88	1.00	1.00	1.00
Predicate	Inference	Existential transitivity	Unrelated	0.88	1.00	1.00	0.88
Predicate	Fallacy	Universal denying a conjunct	Inference	0.88	1.00	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Unrelated	0.88	0.80	0.96	0.92
Predicate	Inference	Universal modus ponens	Inference	0.88	0.96	1.00	1.00
Predicate	Inference	Universal modus tollens	Unrelated	0.88	0.96	1.00	1.00
Predicate	Inference	Universal transitivity	Unrelated	0.88	0.92	1.00	1.00
Propositional	Equivalent	De morgan's laws	Contradiction	0.92	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Contradiction	0.92	1.00	1.00	0.92
Predicate	Equivalent	Law of quantifier movement	Contradiction	0.92	0.84	0.84	0.96
Propositional	Equivalent	Associative laws	Contradiction	0.92	0.92	1.00	1.00
Propositional	Equivalent	Biconditional laws	Unrelated	0.92	0.96	1.00	0.96
Propositional	Inference	Disjunction elimination	Inference	0.92	1.00	1.00	1.00
Predicate	Inference	Existential biconditional elimination	Inference	0.92	0.72	0.56	1.00
Predicate	Equivalent	Existential commutative laws	Unrelated	0.92	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Inference	0.92	0.96	1.00	1.00
Propositional	Inference	Modus tollens	Unrelated	0.92	1.00	1.00	0.96
Predicate	Equivalent	Universal de morgan's laws	Inference	0.92	0.96	1.00	1.00
Predicate	Inference	Universal disjunctive syllogism	Unrelated	0.92	0.92	0.96	0.96
Propositional	Equivalent	De morgan's laws	Inference	0.96	0.96	1.00	0.96
Predicate	Equivalent	Law of quantifier negation	Contradiction	0.96	1.00	1.00	1.00
Propositional	Equivalent	Biconditional laws	Inference	0.96	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Unrelated	0.96	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Unrelated	0.96	1.00	1.00	1.00
Propositional	Equivalent	Distributive laws	Inference	0.96	0.68	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Contradiction	0.96	1.00	1.00	1.00
Predicate	Fallacy	Existential affirming a disjunct	Inference	0.96	1.00	1.00	1.00
Predicate	Fallacy	Existential affirming the consequent	Inference	0.96	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Unrelated	0.96	0.96	1.00	1.00
Predicate	Equivalent	Existential biconditional laws	Contradiction	0.96	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Unrelated	0.96	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Inference	0.96	0.12	0.52	0.80
Predicate	Inference	Existential generalization	Inference	0.96	1.00	1.00	0.96
Predicate	Inference	Existential modus ponens	Unrelated	0.96	0.96	1.00	0.96
Predicate	Inference	Existential modus tollens	Contradiction	0.96	1.00	1.00	0.96
Predicate	Inference	Existential simplification	Inference	0.96	1.00	1.00	1.00
Propositional	Inference	Simplification	Unrelated	0.96	1.00	1.00	1.00
Propositional	Inference	Transitivity	Inference	0.96	1.00	1.00	1.00
Predicate	Fallacy	Universal affirming a disjunct	Inference	0.96	1.00	1.00	0.96
Predicate	Inference	Universal biconditional elimination	Unrelated	0.96	1.00	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Inference	0.96	0.56	1.00	1.00
Predicate	Equivalent	Universal complement laws	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Inference	0.96	0.96	1.00	1.00
Propositional	Equivalent	De morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier movement	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Inference	1.00	0.96	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Affirming a disjunct	Inference	1.00	1.00	1.00	1.00
Propositional	Fallacy	Affirming the consequent	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Associative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Associative laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Contradiction	1.00	1.00	1.00	0.96
Propositional	Inference	Biconditional elimination	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Inference	1.00	1.00	1.00	1.00

Continued on next page

Table 15: Break-down of the accuracy of GPT-4o on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Propositional	Equivalent	Biconditional laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Unrelated	1.00	0.96	1.00	1.00
Propositional	Equivalent	Conditional laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Conditional laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Distributive laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Distributive laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Inference	1.00	0.96	1.00	1.00
Predicate	Inference	Existential addition	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Inference	1.00	0.88	1.00	0.92
Predicate	Equivalent	Existential biconditional laws	Inference	1.00	0.40	1.00	1.00
Predicate	Equivalent	Existential biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Inference	1.00	0.92	1.00	1.00
Predicate	Equivalent	Existential complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential conditional laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Existential illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential resolution	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Fallacy	Illicit major	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Resolution	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Resolution	Unrelated	1.00	0.84	0.92	0.88
Propositional	Inference	Simplification	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Simplification	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Transitivity	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Transitivity	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Undistributed middle	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Contradiction	1.00	1.00	0.96	1.00
Predicate	Equivalent	Universal de morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal affirming the consequent	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Unrelated	1.00	0.96	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Inference	1.00	0.92	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Unrelated	1.00	1.00	1.00	1.00

Continued on next page

Table 15: Break-down of the accuracy of GPT-4o on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Equivalent	Universal complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Inference	1.00	0.92	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Unrelated	1.00	0.96	1.00	1.00
Predicate	Inference	Universal generalization	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus tollens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal resolution	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal resolution	Unrelated	1.00	1.00	0.92	0.76
Predicate	Inference	Universal simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal simplification	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Contradiction	1.00	1.00	1.00	1.00

Table 16: Break-down of the accuracy of Gemini-1.5 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Fallacy	Existential denying the antecedent	Inference	0.00	0.60	0.48	0.84
Predicate	Inference	Existential biconditional introduction	Unrelated	0.08	0.28	0.12	0.32
Predicate	Inference	Existential biconditional introduction	Contradiction	0.16	0.52	0.96	0.88
Predicate	Inference	Existential resolution	Unrelated	0.20	0.40	0.44	0.80
Propositional	Fallacy	Denying the antecedent	Inference	0.32	0.80	0.92	0.72
Predicate	Fallacy	Existential denying a conjunct	Inference	0.32	0.40	0.64	0.88
Predicate	Inference	Universal instantiation	Inference	0.32	1.00	0.64	0.64
Predicate	Inference	Universal disjunctive syllogism	Contradiction	0.44	1.00	1.00	1.00
Predicate	Inference	Existential transitivity	Unrelated	0.48	0.96	0.84	0.68
Predicate	Equivalent	Existential conditional laws	Contradiction	0.52	0.88	0.84	0.72
Predicate	Fallacy	Universal denying the antecedent	Inference	0.52	0.88	0.80	0.76
Predicate	Inference	Universal generalization	Inference	0.52	0.52	0.52	0.52
Predicate	Equivalent	Existential conditional laws	Inference	0.56	0.36	0.84	0.88
Predicate	Inference	Universal resolution	Contradiction	0.56	0.68	0.80	0.88
Predicate	Inference	Universal biconditional introduction	Contradiction	0.64	1.00	1.00	1.00
Predicate	Fallacy	Existential affirming a disjunct	Inference	0.68	0.64	0.44	0.72
Predicate	Equivalent	Universal conditional laws	Contradiction	0.68	0.72	0.88	0.92
Propositional	Equivalent	Conditional laws	Contradiction	0.72	0.64	0.88	0.88
Predicate	Equivalent	Existential biconditional laws	Contradiction	0.72	0.96	0.88	0.88
Predicate	Equivalent	Law of quantifier movement	Inference	0.76	0.48	0.56	0.84
Predicate	Equivalent	Law of quantifier negation	Contradiction	0.76	0.96	0.80	0.88
Propositional	Inference	Disjunctive syllogism	Contradiction	0.76	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Unrelated	0.76	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Unrelated	0.76	0.88	0.96	0.92
Predicate	Fallacy	Existential fallacy	Inference	0.76	0.84	0.80	0.76
Predicate	Inference	Universal transitivity	Contradiction	0.76	1.00	0.96	1.00
Predicate	Equivalent	Law of quantifier movement	Contradiction	0.80	0.76	0.88	0.96
Propositional	Inference	Biconditional introduction	Contradiction	0.80	0.96	1.00	0.92
Predicate	Inference	Existential biconditional elimination	Unrelated	0.80	1.00	1.00	1.00
Propositional	Inference	Resolution	Unrelated	0.80	0.68	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Inference	0.80	0.64	0.88	1.00
Predicate	Inference	Universal generalization	Contradiction	0.80	0.96	1.00	1.00
Propositional	Equivalent	Conditional laws	Inference	0.84	0.64	1.00	1.00
Predicate	Inference	Existential modus tollens	Unrelated	0.84	1.00	0.96	1.00
Predicate	Equivalent	Universal biconditional laws	Contradiction	0.84	1.00	0.96	0.96
Predicate	Equivalent	Universal idempotent laws	Inference	0.84	0.76	1.00	1.00
Predicate	Inference	Universal simplification	Inference	0.84	0.60	1.00	1.00
Predicate	Equivalent	Existential biconditional laws	Unrelated	0.88	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Unrelated	0.88	0.96	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Unrelated	0.88	1.00	1.00	1.00
Propositional	Inference	Transitivity	Contradiction	0.88	1.00	1.00	1.00
Predicate	Fallacy	Universal affirming a disjunct	Inference	0.88	1.00	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Inference	0.88	1.00	1.00	0.96
Predicate	Equivalent	Law of quantifier distribution	Contradiction	0.92	1.00	1.00	0.96
Propositional	Equivalent	Biconditional laws	Contradiction	0.92	0.96	0.96	0.92
Propositional	Inference	Disjunctive syllogism	Inference	0.92	1.00	1.00	1.00

Continued on next page

Table 16: Break-down of the accuracy of Gemini-1.5 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Existential biconditional elimination	Inference	0.92	0.96	0.96	0.88
Predicate	Equivalent	Existential commutative laws	Unrelated	0.92	0.96	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Inference	0.92	0.84	1.00	1.00
Predicate	Inference	Universal conjunction	Inference	0.92	0.96	1.00	1.00
Predicate	Inference	Universal disjunctive syllogism	Inference	0.92	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Inference	0.92	0.96	1.00	1.00
Predicate	Inference	Universal modus tollens	Inference	0.92	0.80	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Inference	0.96	0.96	0.92	0.96
Propositional	Equivalent	Associative laws	Unrelated	0.96	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Unrelated	0.96	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Contradiction	0.96	1.00	1.00	0.96
Predicate	Equivalent	Existential de morgan's laws	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Existential addition	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Existential biconditional introduction	Inference	0.96	0.84	0.96	0.92
Predicate	Inference	Existential disjunctive syllogism	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Existential modus tollens	Inference	0.96	0.92	0.92	0.96
Predicate	Inference	Existential simplification	Unrelated	0.96	1.00	1.00	1.00
Predicate	Fallacy	Illicit minor	Inference	0.96	1.00	1.00	0.96
Propositional	Inference	Resolution	Contradiction	0.96	0.96	1.00	1.00
Propositional	Inference	Resolution	Inference	0.96	0.96	1.00	1.00
Propositional	Inference	Simplification	Unrelated	0.96	0.96	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Inference	0.96	1.00	1.00	1.00
Predicate	Fallacy	Universal denying a conjunct	Inference	0.96	0.96	0.96	0.92
Predicate	Inference	Universal disjunctive syllogism	Unrelated	0.96	0.96	0.96	1.00
Predicate	Inference	Universal simplification	Contradiction	0.96	1.00	1.00	1.00
Propositional	Equivalent	De morgan's laws	Contradiction	1.00	0.80	0.76	0.72
Propositional	Equivalent	De morgan's laws	Inference	1.00	0.92	1.00	1.00
Propositional	Equivalent	De morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier movement	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Inference	1.00	0.88	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Unrelated	1.00	0.96	1.00	1.00
Propositional	Inference	Addition	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Unrelated	1.00	0.96	1.00	1.00
Propositional	Fallacy	Affirming a disjunct	Inference	1.00	1.00	1.00	0.92
Propositional	Fallacy	Affirming the consequent	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Associative laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Associative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Contradiction	1.00	0.96	1.00	1.00
Propositional	Inference	Biconditional elimination	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Inference	1.00	0.92	1.00	1.00
Propositional	Inference	Biconditional introduction	Unrelated	1.00	1.00	0.96	0.88
Propositional	Equivalent	Biconditional laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Contradiction	1.00	0.96	1.00	1.00
Propositional	Equivalent	Complement laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Conditional laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Unrelated	1.00	0.96	1.00	1.00
Propositional	Fallacy	Denying a conjunct	Inference	1.00	0.80	0.96	0.96
Propositional	Inference	Disjunction elimination	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Distributive laws	Contradiction	1.00	0.96	1.00	0.96
Propositional	Equivalent	Distributive laws	Inference	1.00	0.92	1.00	1.00
Propositional	Equivalent	Distributive laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Contradiction	1.00	1.00	0.96	0.92
Predicate	Equivalent	Existential de morgan's laws	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Inference	1.00	0.76	1.00	1.00
Predicate	Fallacy	Existential affirming the consequent	Inference	1.00	1.00	0.96	0.88
Predicate	Equivalent	Existential associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential biconditional elimination	Contradiction	1.00	1.00	1.00	0.96
Predicate	Equivalent	Existential biconditional laws	Inference	1.00	0.88	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Inference	1.00	0.96	1.00	1.00
Predicate	Equivalent	Existential complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Contradiction	1.00	0.96	1.00	1.00

Continued on next page

Table 16: Break-down of the accuracy of Gemini-1.5 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Existential conjunction	Inference	1.00	0.96	0.92	0.96
Predicate	Inference	Existential disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Contradiction	1.00	1.00	0.96	0.96
Predicate	Inference	Existential disjunctive syllogism	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential distributive laws	Inference	1.00	0.80	1.00	0.96
Predicate	Inference	Existential generalization	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Inference	1.00	0.84	1.00	1.00
Predicate	Inference	Existential generalization	Unrelated	1.00	0.96	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Inference	1.00	0.92	1.00	1.00
Predicate	Fallacy	Existential illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Inference	1.00	0.96	0.96	1.00
Predicate	Inference	Existential modus ponens	Unrelated	1.00	1.00	1.00	0.96
Predicate	Inference	Existential modus tollens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential resolution	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential resolution	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential transitivity	Contradiction	1.00	0.76	0.84	0.92
Predicate	Inference	Existential transitivity	Inference	1.00	0.92	0.96	0.96
Propositional	Equivalent	Idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Inference	1.00	0.88	1.00	1.00
Propositional	Equivalent	Idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Fallacy	Illicit major	Inference	1.00	1.00	0.96	0.92
Propositional	Inference	Modus ponens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Inference	1.00	0.96	1.00	1.00
Propositional	Inference	Modus tollens	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Simplification	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Simplification	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Transitivity	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Transitivity	Unrelated	1.00	1.00	0.96	1.00
Predicate	Fallacy	Undistributed middle	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Contradiction	1.00	1.00	0.80	0.84
Predicate	Equivalent	Universal de morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Inference	1.00	0.96	1.00	1.00
Predicate	Inference	Universal addition	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal affirming the consequent	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Inference	1.00	0.84	1.00	0.96
Predicate	Equivalent	Universal associative laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Contradiction	1.00	1.00	0.92	0.96
Predicate	Inference	Universal biconditional introduction	Unrelated	1.00	0.96	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Inference	1.00	0.96	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Inference	1.00	0.92	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Inference	1.00	0.84	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Contradiction	1.00	1.00	0.92	0.96
Predicate	Inference	Universal disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Contradiction	1.00	0.88	0.92	1.00
Predicate	Equivalent	Universal distributive laws	Inference	1.00	0.68	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus tollens	Contradiction	1.00	1.00	0.92	0.96
Predicate	Inference	Universal modus tollens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal resolution	Inference	1.00	0.88	1.00	1.00

Continued on next page

Table 16: Break-down of the accuracy of Gemini-1.5 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Universal resolution	Unrelated	1.00	0.96	1.00	1.00
Predicate	Inference	Universal simplification	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Unrelated	1.00	1.00	1.00	1.00

Table 17: Break-down of the accuracy of Llama3 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Fallacy	Existential denying the antecedent	Inference	0.00	0.72	0.44	0.40
Predicate	Inference	Existential resolution	Unrelated	0.04	0.28	0.16	0.12
Predicate	Inference	Existential biconditional introduction	Unrelated	0.08	0.64	0.20	0.04
Predicate	Fallacy	Universal denying the antecedent	Inference	0.12	0.76	0.76	0.52
Predicate	Fallacy	Existential denying a conjunct	Inference	0.16	0.16	0.60	0.48
Predicate	Fallacy	Universal denying a conjunct	Inference	0.16	0.76	0.52	0.12
Propositional	Fallacy	Denying the antecedent	Inference	0.28	0.88	0.52	0.16
Predicate	Fallacy	Existential fallacy	Inference	0.28	0.80	0.28	0.28
Predicate	Inference	Existential transitivity	Unrelated	0.36	0.76	0.76	0.88
Propositional	Inference	Transitivity	Contradiction	0.40	1.00	0.96	1.00
Predicate	Equivalent	Existential distributive laws	Inference	0.44	0.60	0.64	0.76
Predicate	Equivalent	Existential conditional laws	Contradiction	0.52	0.92	0.68	0.76
Predicate	Inference	Universal generalization	Inference	0.52	0.40	0.52	0.44
Propositional	Equivalent	Conditional laws	Contradiction	0.56	0.76	0.92	1.00
Predicate	Equivalent	Existential biconditional laws	Contradiction	0.60	0.92	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Contradiction	0.60	0.76	1.00	1.00
Predicate	Inference	Universal transitivity	Contradiction	0.60	1.00	1.00	1.00
Propositional	Fallacy	Denying a conjunct	Inference	0.64	0.84	0.60	0.60
Predicate	Equivalent	Existential conditional laws	Inference	0.64	0.48	0.96	1.00
Propositional	Equivalent	De morgan’s laws	Contradiction	0.68	0.88	1.00	0.52
Predicate	Fallacy	Illicit major	Inference	0.68	1.00	0.84	0.60
Predicate	Equivalent	Law of quantifier movement	Inference	0.72	0.48	0.56	1.00
Predicate	Inference	Universal instantiation	Inference	0.72	0.76	0.92	1.00
Predicate	Inference	Existential transitivity	Contradiction	0.76	0.72	1.00	0.96
Predicate	Inference	Universal disjunctive syllogism	Contradiction	0.76	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier movement	Contradiction	0.80	0.84	0.96	0.88
Predicate	Equivalent	Law of quantifier negation	Contradiction	0.80	1.00	0.76	0.96
Propositional	Inference	Disjunctive syllogism	Inference	0.80	0.96	1.00	1.00
Predicate	Inference	Existential conjunction	Unrelated	0.80	0.96	0.96	0.96
Predicate	Inference	Universal resolution	Unrelated	0.80	0.96	0.80	0.56
Propositional	Equivalent	Conditional laws	Inference	0.84	0.96	1.00	1.00
Predicate	Equivalent	Existential associative laws	Unrelated	0.84	0.96	1.00	1.00
Predicate	Inference	Existential biconditional elimination	Contradiction	0.84	1.00	0.92	0.96
Predicate	Equivalent	Existential distributive laws	Unrelated	0.84	1.00	1.00	0.96
Predicate	Fallacy	Illicit minor	Inference	0.84	0.92	0.48	0.20
Propositional	Inference	Resolution	Unrelated	0.84	0.72	0.76	0.72
Predicate	Equivalent	Universal distributive laws	Inference	0.84	0.64	0.64	0.92
Predicate	Equivalent	Existential de morgan’s laws	Contradiction	0.88	1.00	1.00	0.64
Predicate	Fallacy	Existential affirming the consequent	Inference	0.88	1.00	0.80	0.48
Predicate	Inference	Existential biconditional introduction	Inference	0.88	0.92	0.96	1.00
Predicate	Equivalent	Existential commutative laws	Unrelated	0.88	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Contradiction	0.88	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Inference	0.88	0.88	0.96	0.92
Predicate	Fallacy	Universal affirming a disjunct	Inference	0.88	1.00	0.96	0.92
Predicate	Inference	Universal generalization	Contradiction	0.88	0.96	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Contradiction	0.92	1.00	0.96	1.00
Propositional	Equivalent	Associative laws	Contradiction	0.92	0.88	0.92	0.92
Propositional	Equivalent	Associative laws	Unrelated	0.92	1.00	1.00	1.00
Predicate	Fallacy	Existential affirming a disjunct	Inference	0.92	0.84	0.84	0.68
Predicate	Inference	Existential biconditional elimination	Unrelated	0.92	0.96	1.00	0.96
Predicate	Inference	Existential resolution	Contradiction	0.92	0.76	0.96	0.80
Predicate	Inference	Universal biconditional introduction	Contradiction	0.92	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Inference	0.92	0.96	0.88	1.00
Predicate	Equivalent	Universal biconditional laws	Contradiction	0.92	1.00	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Inference	0.92	0.88	0.92	1.00
Predicate	Equivalent	Universal conditional laws	Inference	0.92	0.80	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Unrelated	0.92	1.00	1.00	0.96
Predicate	Inference	Universal resolution	Contradiction	0.92	0.72	0.72	0.60
Predicate	Inference	Universal transitivity	Unrelated	0.92	1.00	1.00	1.00
Propositional	Equivalent	De morgan’s laws	Inference	0.96	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Unrelated	0.96	1.00	1.00	0.96
Propositional	Equivalent	Biconditional laws	Unrelated	0.96	1.00	1.00	0.96
Propositional	Equivalent	Commutative laws	Unrelated	0.96	1.00	1.00	1.00
Propositional	Equivalent	Conditional laws	Unrelated	0.96	0.96	1.00	1.00
Propositional	Equivalent	Distributive laws	Unrelated	0.96	1.00	1.00	0.96
Predicate	Equivalent	Existential de morgan’s laws	Unrelated	0.96	0.96	0.96	1.00
Predicate	Equivalent	Existential biconditional laws	Inference	0.96	0.96	1.00	1.00
Predicate	Equivalent	Existential conditional laws	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Unrelated	0.96	1.00	1.00	1.00

Continued on next page

Table 17: Break-down of the accuracy of Llama3 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Equivalent	Existential distributive laws	Contradiction	0.96	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Unrelated	0.96	0.96	1.00	1.00
Predicate	Inference	Existential modus tollens	Unrelated	0.96	1.00	0.96	0.92
Propositional	Inference	Modus ponens	Inference	0.96	1.00	1.00	1.00
Propositional	Inference	Simplification	Unrelated	0.96	0.96	1.00	1.00
Propositional	Inference	Transitivity	Inference	0.96	1.00	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Contradiction	0.96	0.84	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Unrelated	0.96	0.96	0.96	0.96
Predicate	Inference	Universal disjunctive syllogism	Unrelated	0.96	0.96	0.96	0.96
Predicate	Inference	Universal modus ponens	Inference	0.96	0.96	1.00	1.00
Propositional	Equivalent	De morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Inference	1.00	0.92	0.96	0.88
Predicate	Equivalent	Law of quantifier distribution	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier movement	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Law of quantifier negation	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Addition	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Affirming a disjunct	Inference	1.00	1.00	1.00	1.00
Propositional	Fallacy	Affirming the consequent	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Associative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Unrelated	1.00	1.00	1.00	0.92
Propositional	Equivalent	Biconditional laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Biconditional laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Contradiction	1.00	1.00	0.92	0.92
Propositional	Inference	Conjunction	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Conjunction	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Contradiction	1.00	1.00	0.96	0.92
Propositional	Inference	Disjunctive syllogism	Unrelated	1.00	1.00	1.00	1.00
Propositional	Equivalent	Distributive laws	Contradiction	1.00	1.00	0.92	0.96
Propositional	Equivalent	Distributive laws	Inference	1.00	0.92	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Inference	1.00	0.96	0.96	1.00
Predicate	Inference	Existential biconditional elimination	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential biconditional introduction	Contradiction	1.00	0.76	1.00	1.00
Predicate	Equivalent	Existential biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Inference	1.00	0.96	0.92	0.92
Predicate	Inference	Existential disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Existential illicit commutativity	Inference	1.00	1.00	1.00	0.96
Predicate	Inference	Existential modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Inference	1.00	0.96	1.00	1.00
Predicate	Inference	Existential modus tollens	Contradiction	1.00	0.96	1.00	0.92
Predicate	Inference	Existential modus tollens	Inference	1.00	1.00	0.92	1.00
Predicate	Inference	Existential resolution	Inference	1.00	0.96	1.00	1.00
Predicate	Inference	Existential simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential transitivity	Inference	1.00	0.96	1.00	1.00

Continued on next page

Table 17: Break-down of the accuracy of Llama3 on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Propositional	Equivalent	Idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Illicit commutativity	Inference	1.00	1.00	1.00	0.92
Propositional	Inference	Modus ponens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Unrelated	1.00	0.96	1.00	1.00
Propositional	Inference	Modus tollens	Contradiction	1.00	1.00	0.92	1.00
Propositional	Inference	Modus tollens	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Resolution	Contradiction	1.00	1.00	1.00	0.96
Propositional	Inference	Resolution	Inference	1.00	0.84	1.00	1.00
Propositional	Inference	Simplification	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Simplification	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Transitivity	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Undistributed middle	Inference	1.00	1.00	1.00	0.96
Predicate	Equivalent	Universal de morgan's laws	Inference	1.00	1.00	1.00	0.92
Predicate	Equivalent	Universal de morgan's laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal addition	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal affirming the consequent	Inference	1.00	1.00	1.00	0.80
Predicate	Equivalent	Universal associative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Inference	1.00	0.96	1.00	1.00
Predicate	Equivalent	Universal associative laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional elimination	Inference	1.00	0.88	0.96	1.00
Predicate	Inference	Universal biconditional elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Unrelated	1.00	0.96	1.00	0.96
Predicate	Equivalent	Universal biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Inference	1.00	0.96	1.00	1.00
Predicate	Inference	Universal conjunction	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunction elimination	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunctive syllogism	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal distributive laws	Contradiction	1.00	0.96	1.00	0.88
Predicate	Inference	Universal generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Inference	1.00	0.92	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Fallacy	Universal illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal modus tollens	Contradiction	1.00	0.96	1.00	1.00
Predicate	Inference	Universal modus tollens	Inference	1.00	0.88	1.00	1.00
Predicate	Inference	Universal modus tollens	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal resolution	Inference	1.00	0.76	0.96	1.00
Predicate	Inference	Universal simplification	Contradiction	1.00	0.96	1.00	1.00
Predicate	Inference	Universal simplification	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal simplification	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Inference	1.00	1.00	1.00	1.00

Table 18: Break-down of the accuracy of Mixtral on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Universal disjunction elimination	Inference	0.04	0.16	0.32	0.32
Predicate	Inference	Universal modus ponens	Inference	0.04	0.12	0.08	0.16
Predicate	Equivalent	Existential conditional laws	Inference	0.08	0.20	0.00	0.00
Predicate	Inference	Universal modus tollens	Inference	0.08	0.16	0.12	0.20
Predicate	Equivalent	Universal conditional laws	Inference	0.20	0.16	0.20	0.24
Predicate	Inference	Universal addition	Inference	0.24	0.08	0.04	0.00
Predicate	Inference	Universal disjunctive syllogism	Inference	0.28	0.20	0.20	0.32
Predicate	Inference	Existential addition	Inference	0.32	0.28	0.04	0.00
Propositional	Equivalent	Conditional laws	Inference	0.36	0.40	0.16	0.32
Predicate	Inference	Existential modus ponens	Inference	0.36	0.60	0.56	0.68
Predicate	Inference	Existential transitivity	Inference	0.40	0.36	0.20	0.36
Predicate	Inference	Universal resolution	Inference	0.40	0.76	0.40	0.32
Predicate	Equivalent	Law of quantifier negation	Inference	0.44	0.44	0.72	0.64

Continued on next page

Table 18: Break-down of the accuracy of Mixtral on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Existential biconditional elimination	Inference	0.44	0.60	0.68	0.76
Predicate	Inference	Universal generalization	Inference	0.44	0.44	0.36	0.04
Predicate	Inference	Universal instantiation	Inference	0.44	0.44	0.20	0.24
Predicate	Inference	Universal transitivity	Inference	0.44	0.40	0.56	0.32
Predicate	Inference	Universal biconditional introduction	Contradiction	0.48	0.76	0.80	0.84
Predicate	Equivalent	Law of quantifier movement	Inference	0.52	0.48	0.20	0.16
Propositional	Inference	Biconditional introduction	Contradiction	0.52	0.92	1.00	1.00
Propositional	Inference	Modus ponens	Inference	0.52	0.48	0.88	0.96
Propositional	Inference	Resolution	Unrelated	0.56	0.76	0.84	0.96
Predicate	Inference	Universal biconditional elimination	Inference	0.56	0.56	0.72	0.64
Predicate	Inference	Existential disjunctive syllogism	Inference	0.60	0.48	0.48	0.80
Predicate	Inference	Existential biconditional introduction	Inference	0.64	0.52	0.32	0.44
Predicate	Inference	Existential biconditional introduction	Unrelated	0.64	0.64	0.64	0.68
Predicate	Inference	Existential resolution	Unrelated	0.64	0.64	0.64	0.56
Predicate	Equivalent	Universal de morgan's laws	Inference	0.64	0.60	0.60	0.88
Predicate	Equivalent	Universal biconditional laws	Inference	0.64	0.56	0.96	0.96
Predicate	Inference	Universal conjunction	Inference	0.64	0.68	0.48	0.72
Predicate	Inference	Existential biconditional introduction	Contradiction	0.68	0.36	0.92	1.00
Predicate	Inference	Existential modus tollens	Inference	0.68	0.56	0.44	0.68
Predicate	Inference	Universal biconditional introduction	Inference	0.68	0.64	0.80	0.88
Propositional	Equivalent	Biconditional laws	Contradiction	0.72	0.92	0.92	0.92
Predicate	Equivalent	Existential biconditional laws	Inference	0.72	0.64	0.48	0.72
Propositional	Inference	Modus tollens	Inference	0.72	0.80	0.28	0.64
Propositional	Inference	Transitivity	Inference	0.72	0.68	0.48	0.36
Predicate	Inference	Universal transitivity	Contradiction	0.72	0.48	0.84	1.00
Predicate	Equivalent	Law of quantifier distribution	Inference	0.76	0.76	0.60	0.80
Predicate	Fallacy	Existential affirming a disjunct	Inference	0.76	0.88	0.84	0.84
Predicate	Inference	Existential conjunction	Inference	0.76	0.80	0.44	0.40
Predicate	Inference	Existential transitivity	Unrelated	0.76	0.96	0.88	0.88
Propositional	Inference	Transitivity	Contradiction	0.76	0.52	0.80	0.88
Predicate	Inference	Universal simplification	Inference	0.76	0.76	0.52	0.36
Predicate	Equivalent	Law of quantifier movement	Contradiction	0.80	0.72	0.92	1.00
Propositional	Inference	Disjunction elimination	Inference	0.80	0.88	0.80	0.76
Propositional	Inference	Disjunctive syllogism	Inference	0.80	0.84	0.64	0.64
Predicate	Inference	Existential conjunction	Unrelated	0.80	0.96	0.84	0.96
Predicate	Inference	Universal biconditional elimination	Contradiction	0.80	0.92	0.88	0.88
Predicate	Equivalent	Universal conditional laws	Contradiction	0.80	0.80	0.76	0.84
Predicate	Inference	Universal disjunctive syllogism	Contradiction	0.80	0.84	0.72	0.92
Predicate	Equivalent	Universal distributive laws	Inference	0.80	0.76	0.16	0.16
Predicate	Equivalent	Universal idempotent laws	Inference	0.80	0.84	0.28	0.68
Propositional	Inference	Addition	Inference	0.84	0.76	0.96	0.32
Propositional	Inference	Biconditional elimination	Inference	0.84	0.88	0.56	0.80
Propositional	Inference	Biconditional introduction	Unrelated	0.84	0.96	0.32	0.08
Propositional	Equivalent	Biconditional laws	Inference	0.84	0.88	0.96	1.00
Propositional	Fallacy	Denying the antecedent	Inference	0.84	0.88	0.76	0.44
Predicate	Equivalent	Existential distributive laws	Inference	0.84	0.76	0.24	0.48
Predicate	Inference	Existential transitivity	Contradiction	0.84	0.52	0.96	1.00
Propositional	Inference	Resolution	Inference	0.84	0.96	0.80	0.64
Propositional	Inference	Simplification	Inference	0.84	0.80	1.00	0.84
Predicate	Equivalent	Universal biconditional laws	Contradiction	0.84	0.88	0.92	0.92
Predicate	Equivalent	Law of quantifier distribution	Contradiction	0.88	0.88	0.92	1.00
Propositional	Fallacy	Affirming a disjunct	Inference	0.88	0.96	0.64	0.88
Propositional	Fallacy	Denying a conjunct	Inference	0.88	0.92	0.92	0.84
Predicate	Equivalent	Existential de morgan's laws	Inference	0.88	0.80	0.72	0.88
Predicate	Equivalent	Existential biconditional laws	Contradiction	0.88	0.36	0.88	1.00
Predicate	Equivalent	Existential conditional laws	Contradiction	0.88	0.76	0.84	1.00
Predicate	Inference	Existential generalization	Inference	0.88	0.88	0.68	0.08
Predicate	Equivalent	Universal complement laws	Inference	0.88	0.80	0.48	0.80
Predicate	Inference	Universal generalization	Contradiction	0.88	0.96	0.64	0.96
Predicate	Inference	Universal modus ponens	Contradiction	0.88	1.00	0.96	0.96
Propositional	Inference	Biconditional elimination	Contradiction	0.92	0.88	0.80	0.96
Propositional	Equivalent	Distributive laws	Inference	0.92	0.96	0.76	0.84
Predicate	Equivalent	Existential associative laws	Inference	0.92	0.88	0.84	0.72
Predicate	Inference	Existential biconditional elimination	Contradiction	0.92	0.92	0.96	0.92
Predicate	Inference	Existential biconditional elimination	Unrelated	0.92	1.00	0.96	0.92
Predicate	Equivalent	Existential commutative laws	Inference	0.92	0.88	0.84	0.68
Predicate	Fallacy	Existential denying a conjunct	Inference	0.92	0.92	0.76	0.80
Predicate	Fallacy	Existential denying the antecedent	Inference	0.92	0.88	0.56	0.64
Predicate	Inference	Existential disjunction elimination	Inference	0.92	0.80	0.72	0.72
Predicate	Fallacy	Existential fallacy	Inference	0.92	0.84	0.68	0.84
Predicate	Inference	Existential resolution	Inference	0.92	0.64	0.60	0.56
Predicate	Equivalent	Universal associative laws	Inference	0.92	0.88	0.48	0.64
Predicate	Fallacy	Universal denying the antecedent	Inference	0.92	0.92	0.92	0.80
Propositional	Equivalent	De morgan's laws	Inference	0.96	0.92	1.00	0.96
Predicate	Equivalent	Law of quantifier negation	Contradiction	0.96	0.96	0.76	0.72
Propositional	Inference	Biconditional introduction	Inference	0.96	0.92	0.92	1.00
Propositional	Equivalent	Biconditional laws	Unrelated	0.96	1.00	1.00	1.00
Propositional	Equivalent	Commutative laws	Inference	0.96	0.96	0.76	0.92
Propositional	Inference	Conjunction	Contradiction	0.96	0.96	0.96	0.92

Continued on next page

Table 18: Break-down of the accuracy of Mixtral on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Propositional	Inference	Conjunction	Inference	0.96	0.96	0.92	0.88
Predicate	Inference	Existential disjunctive syllogism	Unrelated	0.96	1.00	0.76	0.96
Predicate	Equivalent	Existential distributive laws	Contradiction	0.96	1.00	0.88	1.00
Predicate	Equivalent	Existential distributive laws	Unrelated	0.96	1.00	0.80	0.96
Predicate	Inference	Existential modus tollens	Unrelated	0.96	1.00	0.96	0.96
Predicate	Inference	Existential simplification	Inference	0.96	0.96	0.60	0.80
Predicate	Fallacy	Undistributed middle	Inference	0.96	1.00	0.84	0.88
Predicate	Equivalent	Universal de morgan's laws	Contradiction	0.96	0.88	0.76	0.92
Predicate	Equivalent	Universal commutative laws	Inference	0.96	0.96	0.80	0.88
Predicate	Inference	Universal disjunction elimination	Contradiction	0.96	1.00	0.96	1.00
Predicate	Inference	Universal resolution	Contradiction	0.96	0.96	0.88	0.92
Propositional	Equivalent	De morgan's laws	Contradiction	1.00	1.00	1.00	0.88
Propositional	Equivalent	De morgan's laws	Unrelated	1.00	1.00	0.84	1.00
Predicate	Equivalent	Law of quantifier distribution	Unrelated	1.00	1.00	0.80	0.84
Predicate	Equivalent	Law of quantifier movement	Unrelated	1.00	1.00	0.80	0.96
Predicate	Equivalent	Law of quantifier negation	Unrelated	1.00	1.00	0.72	0.76
Propositional	Inference	Addition	Contradiction	1.00	1.00	0.72	1.00
Propositional	Inference	Addition	Unrelated	1.00	1.00	1.00	1.00
Propositional	Fallacy	Affirming the consequent	Inference	1.00	1.00	0.96	0.92
Propositional	Equivalent	Associative laws	Contradiction	1.00	1.00	1.00	0.88
Propositional	Equivalent	Associative laws	Inference	1.00	0.96	0.92	0.88
Propositional	Equivalent	Associative laws	Unrelated	1.00	1.00	0.88	1.00
Propositional	Inference	Biconditional elimination	Unrelated	1.00	1.00	0.92	1.00
Propositional	Equivalent	Commutative laws	Contradiction	1.00	1.00	0.96	0.96
Propositional	Equivalent	Commutative laws	Unrelated	1.00	1.00	0.84	0.96
Propositional	Equivalent	Complement laws	Contradiction	1.00	1.00	0.72	1.00
Propositional	Equivalent	Complement laws	Inference	1.00	1.00	0.60	0.88
Propositional	Equivalent	Complement laws	Unrelated	1.00	1.00	0.88	1.00
Propositional	Equivalent	Conditional laws	Contradiction	1.00	0.96	0.92	0.96
Propositional	Equivalent	Conditional laws	Unrelated	1.00	1.00	0.88	1.00
Propositional	Inference	Conjunction	Unrelated	1.00	1.00	0.96	0.96
Propositional	Inference	Disjunction elimination	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunctive syllogism	Contradiction	1.00	0.92	0.92	0.96
Propositional	Inference	Disjunctive syllogism	Unrelated	1.00	1.00	0.96	1.00
Propositional	Equivalent	Distributive laws	Contradiction	1.00	0.96	0.72	0.88
Propositional	Equivalent	Distributive laws	Unrelated	1.00	1.00	0.88	1.00
Predicate	Equivalent	Existential de morgan's laws	Contradiction	1.00	1.00	0.92	0.96
Predicate	Equivalent	Existential de morgan's laws	Unrelated	1.00	1.00	0.84	0.92
Predicate	Inference	Existential addition	Contradiction	1.00	1.00	0.80	0.84
Predicate	Inference	Existential addition	Unrelated	1.00	1.00	0.80	0.92
Predicate	Fallacy	Existential affirming the consequent	Inference	1.00	1.00	0.96	1.00
Predicate	Equivalent	Existential associative laws	Contradiction	1.00	1.00	0.80	1.00
Predicate	Equivalent	Existential associative laws	Unrelated	1.00	1.00	0.96	0.96
Predicate	Equivalent	Existential biconditional laws	Unrelated	1.00	1.00	0.96	1.00
Predicate	Equivalent	Existential commutative laws	Contradiction	1.00	1.00	0.72	1.00
Predicate	Equivalent	Existential commutative laws	Unrelated	1.00	1.00	0.88	0.88
Predicate	Equivalent	Existential complement laws	Contradiction	1.00	1.00	0.76	0.96
Predicate	Equivalent	Existential complement laws	Inference	1.00	0.92	0.72	0.84
Predicate	Equivalent	Existential complement laws	Unrelated	1.00	0.96	0.88	0.88
Predicate	Equivalent	Existential conditional laws	Unrelated	1.00	1.00	0.92	0.96
Predicate	Inference	Existential conjunction	Contradiction	1.00	0.96	0.96	1.00
Predicate	Inference	Existential disjunction elimination	Contradiction	1.00	1.00	0.92	1.00
Predicate	Inference	Existential disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Contradiction	1.00	1.00	0.92	0.96
Predicate	Inference	Existential generalization	Contradiction	1.00	1.00	0.52	0.80
Predicate	Inference	Existential generalization	Unrelated	1.00	1.00	0.88	0.88
Predicate	Equivalent	Existential idempotent laws	Contradiction	1.00	1.00	0.80	0.92
Predicate	Equivalent	Existential idempotent laws	Inference	1.00	1.00	0.52	0.84
Predicate	Equivalent	Existential idempotent laws	Unrelated	1.00	1.00	0.92	0.96
Predicate	Fallacy	Existential illicit commutativity	Inference	1.00	0.96	0.96	1.00
Predicate	Inference	Existential modus ponens	Contradiction	1.00	1.00	0.92	1.00
Predicate	Inference	Existential modus ponens	Unrelated	1.00	1.00	0.96	1.00
Predicate	Inference	Existential modus tollens	Contradiction	1.00	1.00	0.80	1.00
Predicate	Inference	Existential resolution	Contradiction	1.00	1.00	0.92	1.00
Predicate	Inference	Existential simplification	Contradiction	1.00	1.00	0.64	0.68
Predicate	Inference	Existential simplification	Unrelated	1.00	1.00	0.96	0.96
Propositional	Equivalent	Idempotent laws	Contradiction	1.00	1.00	0.88	0.80
Propositional	Equivalent	Idempotent laws	Inference	1.00	0.96	0.60	0.96
Propositional	Equivalent	Idempotent laws	Unrelated	1.00	1.00	0.96	0.96
Propositional	Fallacy	Illicit commutativity	Inference	1.00	1.00	0.92	1.00
Predicate	Fallacy	Illicit major	Inference	1.00	1.00	0.96	0.96
Predicate	Fallacy	Illicit minor	Inference	1.00	0.96	0.96	1.00
Propositional	Inference	Modus ponens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Unrelated	1.00	1.00	0.96	1.00
Propositional	Inference	Modus tollens	Contradiction	1.00	1.00	0.92	0.96
Propositional	Inference	Modus tollens	Unrelated	1.00	1.00	0.84	1.00
Propositional	Inference	Resolution	Contradiction	1.00	1.00	0.96	1.00
Propositional	Inference	Simplification	Contradiction	1.00	1.00	1.00	0.96

Continued on next page

Table 18: Break-down of the accuracy of Mixtral on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Propositional	Inference	Simplification	Unrelated	1.00	1.00	0.92	0.96
Propositional	Inference	Transitivity	Unrelated	1.00	1.00	0.88	0.88
Predicate	Equivalent	Universal de morgan's laws	Unrelated	1.00	1.00	0.84	0.96
Predicate	Inference	Universal addition	Contradiction	1.00	0.96	0.88	0.88
Predicate	Inference	Universal addition	Unrelated	1.00	1.00	0.80	0.72
Predicate	Fallacy	Universal affirming a disjunct	Inference	1.00	1.00	0.88	0.96
Predicate	Fallacy	Universal affirming the consequent	Inference	1.00	0.96	0.92	1.00
Predicate	Equivalent	Universal associative laws	Contradiction	1.00	1.00	0.96	1.00
Predicate	Equivalent	Universal associative laws	Unrelated	1.00	1.00	0.72	0.88
Predicate	Inference	Universal biconditional elimination	Unrelated	1.00	0.88	0.80	0.96
Predicate	Inference	Universal biconditional introduction	Unrelated	1.00	1.00	0.88	0.92
Predicate	Equivalent	Universal biconditional laws	Unrelated	1.00	1.00	0.80	0.92
Predicate	Equivalent	Universal commutative laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Unrelated	1.00	1.00	0.76	0.84
Predicate	Equivalent	Universal complement laws	Contradiction	1.00	1.00	0.96	0.88
Predicate	Equivalent	Universal complement laws	Unrelated	1.00	1.00	0.76	0.80
Predicate	Equivalent	Universal conditional laws	Unrelated	1.00	0.96	0.72	0.96
Predicate	Inference	Universal conjunction	Contradiction	1.00	1.00	0.80	1.00
Predicate	Inference	Universal conjunction	Unrelated	1.00	0.88	0.64	0.68
Predicate	Fallacy	Universal denying a conjunct	Inference	1.00	1.00	0.96	0.96
Predicate	Inference	Universal disjunction elimination	Unrelated	1.00	1.00	1.00	1.00
Predicate	Inference	Universal disjunctive syllogism	Unrelated	1.00	0.96	0.92	1.00
Predicate	Equivalent	Universal distributive laws	Contradiction	1.00	0.96	0.96	1.00
Predicate	Equivalent	Universal distributive laws	Unrelated	1.00	1.00	0.88	0.96
Predicate	Inference	Universal generalization	Unrelated	1.00	1.00	0.80	0.80
Predicate	Equivalent	Universal idempotent laws	Contradiction	1.00	0.92	1.00	0.88
Predicate	Equivalent	Universal idempotent laws	Unrelated	1.00	0.84	0.88	0.84
Predicate	Fallacy	Universal illicit commutativity	Inference	1.00	1.00	0.92	1.00
Predicate	Inference	Universal instantiation	Contradiction	1.00	0.96	0.68	0.84
Predicate	Inference	Universal instantiation	Unrelated	1.00	1.00	0.44	0.84
Predicate	Inference	Universal modus ponens	Unrelated	1.00	0.96	0.84	1.00
Predicate	Inference	Universal modus tollens	Contradiction	1.00	1.00	0.96	1.00
Predicate	Inference	Universal modus tollens	Unrelated	1.00	1.00	0.96	0.92
Predicate	Inference	Universal resolution	Unrelated	1.00	1.00	0.88	0.96
Predicate	Inference	Universal simplification	Contradiction	1.00	1.00	0.80	0.84
Predicate	Inference	Universal simplification	Unrelated	1.00	0.96	0.72	0.84
Predicate	Inference	Universal transitivity	Unrelated	1.00	1.00	0.88	0.96

Table 19: Break-down of the accuracy of ChatGPT on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Existential resolution	Unrelated	0.00	0.40	0.72	0.52
Predicate	Equivalent	Universal de morgan's laws	Contradiction	0.04	0.40	0.24	0.12
Predicate	Equivalent	Existential de morgan's laws	Contradiction	0.08	0.96	0.64	1.00
Predicate	Inference	Existential biconditional introduction	Unrelated	0.08	0.52	0.40	0.52
Predicate	Inference	Universal instantiation	Inference	0.08	0.44	0.04	0.12
Predicate	Fallacy	Existential denying the antecedent	Inference	0.12	0.88	0.76	0.68
Predicate	Inference	Universal disjunctive syllogism	Contradiction	0.12	0.60	0.20	0.04
Predicate	Inference	Universal transitivity	Contradiction	0.12	0.32	1.00	0.68
Propositional	Equivalent	De morgan's laws	Contradiction	0.20	0.80	0.56	0.56
Predicate	Fallacy	Existential denying a conjunct	Inference	0.20	0.60	0.76	0.76
Predicate	Equivalent	Universal distributive laws	Inference	0.24	0.04	0.24	0.20
Predicate	Inference	Universal transitivity	Inference	0.24	0.00	0.44	1.00
Propositional	Inference	Biconditional introduction	Unrelated	0.28	0.72	0.00	0.64
Predicate	Equivalent	Existential de morgan's laws	Unrelated	0.28	0.92	1.00	1.00
Predicate	Fallacy	Existential affirming a disjunct	Inference	0.28	0.72	0.84	0.48
Predicate	Equivalent	Universal conditional laws	Inference	0.28	0.08	0.28	0.52
Predicate	Inference	Universal disjunction elimination	Inference	0.28	0.00	0.72	0.96
Propositional	Equivalent	Conditional laws	Inference	0.32	0.16	0.12	0.44
Predicate	Inference	Existential transitivity	Unrelated	0.32	0.60	0.92	0.24
Predicate	Inference	Universal conjunction	Inference	0.32	0.24	0.64	0.48
Propositional	Inference	Resolution	Unrelated	0.36	0.72	0.96	0.64
Predicate	Fallacy	Universal denying the antecedent	Inference	0.36	0.96	0.72	0.32
Predicate	Inference	Universal generalization	Inference	0.36	0.28	0.00	0.04
Predicate	Equivalent	Law of quantifier movement	Inference	0.40	0.32	0.00	0.20
Propositional	Equivalent	Biconditional laws	Contradiction	0.40	0.68	0.68	0.04
Propositional	Inference	Transitivity	Inference	0.40	0.16	0.80	1.00
Predicate	Inference	Universal resolution	Contradiction	0.40	0.92	0.96	0.40
Predicate	Equivalent	Existential conditional laws	Inference	0.44	0.04	0.16	0.44
Propositional	Inference	Conjunction	Inference	0.48	0.56	0.84	0.84
Predicate	Fallacy	Existential fallacy	Inference	0.48	0.80	0.76	0.60
Predicate	Inference	Universal modus tollens	Inference	0.48	0.16	0.48	0.52
Propositional	Inference	Addition	Inference	0.52	0.16	0.08	0.12
Predicate	Equivalent	Existential biconditional laws	Contradiction	0.52	0.92	0.68	0.64
Predicate	Equivalent	Existential conditional laws	Unrelated	0.52	0.92	1.00	1.00
Predicate	Inference	Existential disjunctive syllogism	Unrelated	0.52	0.88	0.92	1.00
Predicate	Inference	Existential modus tollens	Unrelated	0.52	1.00	0.96	1.00

Continued on next page

Table 19: Break-down of the accuracy of ChatGPT on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Inference	Existential transitivity	Inference	0.52	0.12	0.32	0.80
Propositional	Inference	Modus tollens	Inference	0.52	0.44	0.84	0.96
Propositional	Inference	Transitivity	Unrelated	0.52	0.68	0.76	0.44
Predicate	Inference	Universal addition	Inference	0.52	0.08	0.00	0.00
Predicate	Inference	Universal disjunctive syllogism	Inference	0.52	0.24	0.48	0.56
Propositional	Equivalent	Associative laws	Unrelated	0.56	0.96	1.00	1.00
Predicate	Inference	Existential biconditional elimination	Unrelated	0.56	0.92	0.96	1.00
Predicate	Equivalent	Existential conditional laws	Contradiction	0.56	0.80	1.00	1.00
Predicate	Inference	Universal generalization	Contradiction	0.56	0.88	1.00	1.00
Predicate	Inference	Universal modus ponens	Inference	0.56	0.08	0.60	0.88
Propositional	Inference	Disjunctive syllogism	Unrelated	0.60	1.00	1.00	1.00
Predicate	Equivalent	Existential associative laws	Unrelated	0.60	0.96	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Unrelated	0.60	0.96	1.00	1.00
Predicate	Inference	Existential modus ponens	Unrelated	0.60	0.96	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Contradiction	0.60	0.84	1.00	0.64
Predicate	Fallacy	Universal denying a conjunct	Inference	0.60	0.96	1.00	0.72
Propositional	Equivalent	Conditional laws	Contradiction	0.64	0.76	1.00	0.68
Propositional	Fallacy	Denying a conjunct	Inference	0.64	1.00	1.00	0.76
Predicate	Inference	Existential conjunction	Contradiction	0.64	1.00	1.00	1.00
Predicate	Inference	Existential conjunction	Unrelated	0.64	0.92	1.00	1.00
Predicate	Fallacy	Illicit major	Inference	0.64	0.84	0.96	0.60
Predicate	Equivalent	Universal biconditional laws	Contradiction	0.64	0.96	0.48	0.24
Propositional	Inference	Disjunctive syllogism	Inference	0.68	0.48	0.88	0.96
Predicate	Equivalent	Existential biconditional laws	Unrelated	0.68	0.96	1.00	1.00
Predicate	Inference	Existential simplification	Unrelated	0.68	0.84	1.00	1.00
Predicate	Inference	Universal disjunctive syllogism	Unrelated	0.68	0.96	0.96	0.92
Predicate	Inference	Universal simplification	Inference	0.68	0.44	0.04	0.12
Propositional	Equivalent	De morgan's laws	Unrelated	0.72	0.92	1.00	1.00
Propositional	Inference	Biconditional elimination	Inference	0.72	0.60	0.84	1.00
Propositional	Equivalent	Distributive laws	Unrelated	0.72	0.96	1.00	1.00
Predicate	Inference	Existential addition	Unrelated	0.72	0.96	1.00	1.00
Propositional	Inference	Transitivity	Contradiction	0.72	0.92	1.00	1.00
Predicate	Equivalent	Universal associative laws	Inference	0.72	0.44	0.96	0.64
Predicate	Inference	Universal biconditional elimination	Contradiction	0.72	0.72	0.68	0.36
Predicate	Inference	Universal biconditional elimination	Inference	0.72	0.24	0.60	0.92
Propositional	Fallacy	Affirming a disjunct	Inference	0.76	0.92	0.84	0.48
Propositional	Equivalent	Biconditional laws	Unrelated	0.76	1.00	0.96	0.92
Propositional	Equivalent	Distributive laws	Inference	0.76	0.32	0.20	0.60
Predicate	Inference	Existential conjunction	Inference	0.76	0.72	0.60	0.52
Predicate	Equivalent	Existential distributive laws	Unrelated	0.76	0.88	1.00	1.00
Propositional	Equivalent	Idempotent laws	Unrelated	0.76	1.00	0.96	0.96
Predicate	Inference	Universal biconditional elimination	Unrelated	0.76	1.00	1.00	0.96
Predicate	Equivalent	Universal distributive laws	Contradiction	0.76	0.96	1.00	0.68
Predicate	Inference	Universal modus tollens	Unrelated	0.76	0.96	1.00	1.00
Propositional	Equivalent	Associative laws	Inference	0.80	0.92	1.00	1.00
Propositional	Equivalent	Commutative laws	Unrelated	0.80	1.00	1.00	0.96
Propositional	Inference	Disjunctive syllogism	Contradiction	0.80	0.96	0.28	0.00
Predicate	Inference	Existential biconditional elimination	Inference	0.80	0.32	0.56	0.72
Predicate	Equivalent	Existential complement laws	Unrelated	0.80	0.96	1.00	1.00
Propositional	Inference	Simplification	Unrelated	0.80	0.96	0.96	0.96
Predicate	Inference	Universal biconditional introduction	Unrelated	0.80	0.92	0.68	0.96
Predicate	Inference	Universal disjunction elimination	Contradiction	0.80	0.48	1.00	0.52
Predicate	Equivalent	Law of quantifier distribution	Unrelated	0.84	1.00	1.00	1.00
Propositional	Equivalent	Conditional laws	Unrelated	0.84	0.88	0.96	0.96
Propositional	Inference	Conjunction	Unrelated	0.84	1.00	1.00	1.00
Propositional	Fallacy	Denying the antecedent	Inference	0.84	0.80	0.48	0.08
Predicate	Inference	Existential disjunctive syllogism	Contradiction	0.84	0.88	0.96	1.00
Predicate	Inference	Existential generalization	Inference	0.84	0.60	0.28	0.40
Predicate	Equivalent	Existential idempotent laws	Unrelated	0.84	0.96	1.00	1.00
Propositional	Inference	Modus ponens	Unrelated	0.84	1.00	1.00	1.00
Propositional	Inference	Modus tollens	Unrelated	0.84	1.00	1.00	1.00
Predicate	Inference	Universal conjunction	Contradiction	0.84	0.80	1.00	0.96
Predicate	Inference	Universal conjunction	Unrelated	0.84	1.00	1.00	1.00
Predicate	Fallacy	Existential affirming the consequent	Inference	0.88	0.88	1.00	0.88
Predicate	Equivalent	Existential biconditional laws	Inference	0.88	0.40	0.88	0.92
Predicate	Inference	Existential transitivity	Contradiction	0.88	0.96	1.00	0.96
Predicate	Equivalent	Universal de morgan's laws	Unrelated	0.88	1.00	1.00	1.00
Predicate	Inference	Universal addition	Unrelated	0.88	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Unrelated	0.88	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Unrelated	0.88	0.96	1.00	0.96
Predicate	Inference	Universal modus tollens	Contradiction	0.88	1.00	1.00	0.92
Predicate	Inference	Universal resolution	Unrelated	0.88	1.00	1.00	0.88
Predicate	Equivalent	Law of quantifier distribution	Contradiction	0.92	0.92	0.80	0.72
Predicate	Equivalent	Law of quantifier movement	Contradiction	0.92	0.80	0.96	1.00
Predicate	Equivalent	Law of quantifier movement	Unrelated	0.92	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Contradiction	0.92	0.84	0.08	0.28
Propositional	Equivalent	Commutative laws	Inference	0.92	0.92	1.00	1.00
Predicate	Equivalent	Existential associative laws	Contradiction	0.92	0.96	1.00	0.96
Predicate	Equivalent	Existential distributive laws	Inference	0.92	0.24	0.40	0.40

Continued on next page

Table 19: Break-down of the accuracy of ChatGPT on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Propositional	Inference	Modus ponens	Inference	0.92	0.80	1.00	0.84
Predicate	Equivalent	Universal associative laws	Contradiction	0.92	1.00	1.00	0.84
Predicate	Inference	Universal biconditional introduction	Contradiction	0.92	0.96	0.60	0.84
Predicate	Equivalent	Universal distributive laws	Unrelated	0.92	1.00	1.00	1.00
Predicate	Inference	Universal instantiation	Contradiction	0.92	1.00	1.00	1.00
Predicate	Inference	Universal simplification	Unrelated	0.92	0.96	1.00	1.00
Predicate	Equivalent	Law of quantifier distribution	Inference	0.96	0.76	0.72	0.80
Predicate	Equivalent	Law of quantifier negation	Unrelated	0.96	0.96	1.00	1.00
Propositional	Inference	Addition	Unrelated	0.96	1.00	1.00	1.00
Propositional	Fallacy	Affirming the consequent	Inference	0.96	0.88	0.96	0.92
Propositional	Equivalent	Associative laws	Contradiction	0.96	1.00	0.96	0.92
Propositional	Inference	Biconditional elimination	Contradiction	0.96	1.00	0.92	0.88
Propositional	Inference	Biconditional elimination	Unrelated	0.96	1.00	1.00	1.00
Propositional	Inference	Biconditional introduction	Inference	0.96	0.44	1.00	1.00
Propositional	Equivalent	Biconditional laws	Inference	0.96	0.60	1.00	1.00
Propositional	Equivalent	Commutative laws	Contradiction	0.96	0.76	0.96	0.60
Predicate	Equivalent	Existential associative laws	Inference	0.96	0.88	0.96	0.84
Predicate	Inference	Existential biconditional introduction	Inference	0.96	0.40	0.96	1.00
Predicate	Equivalent	Existential commutative laws	Contradiction	0.96	1.00	1.00	0.96
Predicate	Inference	Existential modus ponens	Inference	0.96	0.28	0.76	0.96
Predicate	Inference	Existential modus tollens	Inference	0.96	0.32	0.32	0.32
Predicate	Fallacy	Illicit minor	Inference	0.96	1.00	1.00	1.00
Predicate	Inference	Universal addition	Contradiction	0.96	1.00	1.00	1.00
Predicate	Equivalent	Universal associative laws	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Universal biconditional introduction	Inference	0.96	0.52	1.00	1.00
Predicate	Equivalent	Universal biconditional laws	Inference	0.96	0.36	0.84	0.96
Predicate	Inference	Universal disjunction elimination	Unrelated	0.96	1.00	0.96	0.80
Predicate	Inference	Universal generalization	Unrelated	0.96	0.96	1.00	1.00
Predicate	Inference	Universal instantiation	Unrelated	0.96	1.00	1.00	1.00
Predicate	Inference	Universal modus ponens	Contradiction	0.96	0.88	1.00	0.80
Propositional	Equivalent	De morgan's laws	Inference	1.00	0.16	0.52	1.00
Predicate	Equivalent	Law of quantifier negation	Contradiction	1.00	1.00	0.88	0.80
Predicate	Equivalent	Law of quantifier negation	Inference	1.00	0.48	0.48	0.68
Propositional	Inference	Addition	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Inference	1.00	1.00	1.00	1.00
Propositional	Equivalent	Complement laws	Unrelated	1.00	0.92	1.00	1.00
Propositional	Inference	Conjunction	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Disjunction elimination	Contradiction	1.00	1.00	1.00	0.20
Propositional	Inference	Disjunction elimination	Inference	1.00	0.60	0.92	1.00
Propositional	Inference	Disjunction elimination	Unrelated	1.00	1.00	1.00	0.96
Propositional	Equivalent	Distributive laws	Contradiction	1.00	0.84	1.00	1.00
Predicate	Equivalent	Existential de morgan's laws	Inference	1.00	0.76	0.24	0.80
Predicate	Inference	Existential addition	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential addition	Inference	1.00	0.12	0.20	0.00
Predicate	Inference	Existential biconditional elimination	Contradiction	1.00	1.00	1.00	0.96
Predicate	Inference	Existential biconditional introduction	Contradiction	1.00	0.84	1.00	1.00
Predicate	Equivalent	Existential commutative laws	Inference	1.00	0.64	0.92	0.88
Predicate	Equivalent	Existential complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential disjunction elimination	Contradiction	1.00	1.00	1.00	0.84
Predicate	Inference	Existential disjunction elimination	Inference	1.00	0.72	0.68	1.00
Predicate	Inference	Existential disjunction elimination	Unrelated	1.00	1.00	1.00	0.96
Predicate	Inference	Existential disjunctive syllogism	Inference	1.00	0.36	0.48	0.44
Predicate	Equivalent	Existential distributive laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential generalization	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Existential idempotent laws	Inference	1.00	1.00	0.88	0.88
Predicate	Fallacy	Existential illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus ponens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential modus tollens	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential resolution	Contradiction	1.00	0.96	1.00	1.00
Predicate	Inference	Existential resolution	Inference	1.00	0.64	0.52	0.84
Predicate	Inference	Existential simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Existential simplification	Inference	1.00	0.88	0.96	0.68
Propositional	Equivalent	Idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Propositional	Equivalent	Idempotent laws	Inference	1.00	1.00	0.56	0.52
Propositional	Fallacy	Illicit commutativity	Inference	1.00	1.00	1.00	1.00
Propositional	Inference	Modus ponens	Contradiction	1.00	1.00	1.00	0.76
Propositional	Inference	Modus tollens	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Resolution	Contradiction	1.00	0.72	1.00	0.92
Propositional	Inference	Resolution	Inference	1.00	0.60	1.00	1.00
Propositional	Inference	Simplification	Contradiction	1.00	1.00	1.00	1.00
Propositional	Inference	Simplification	Inference	1.00	1.00	1.00	0.92
Predicate	Fallacy	Undistributed middle	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal de morgan's laws	Inference	1.00	0.32	0.68	0.88
Predicate	Fallacy	Universal affirming a disjunct	Inference	1.00	1.00	1.00	0.92
Predicate	Fallacy	Universal affirming the consequent	Inference	1.00	1.00	1.00	0.96

Continued on next page

Table 19: Break-down of the accuracy of ChatGPT on all rules (sorted by zero-shot accuracy).

Logic	Rule category	Rule	Problem	Zero shot	Zero shot cot	Random icl	Weak
Predicate	Equivalent	Universal biconditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Contradiction	1.00	1.00	1.00	0.96
Predicate	Equivalent	Universal commutative laws	Inference	1.00	0.56	1.00	1.00
Predicate	Equivalent	Universal commutative laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Inference	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal complement laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal conditional laws	Unrelated	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Contradiction	1.00	1.00	1.00	1.00
Predicate	Equivalent	Universal idempotent laws	Inference	1.00	0.88	0.68	0.80
Predicate	Fallacy	Universal illicit commutativity	Inference	1.00	1.00	1.00	1.00
Predicate	Inference	Universal resolution	Inference	1.00	0.16	0.24	0.88
Predicate	Inference	Universal simplification	Contradiction	1.00	1.00	1.00	1.00
Predicate	Inference	Universal transitivity	Unrelated	1.00	1.00	0.96	0.36