Comparing Inferential Strategies of Humans and Large Language Models in Deductive Reasoning

Philipp Mondorf[1, 2] and **Barbara Plank**[1, 2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
`p.mondorf@lmu.de` `b.plank@lmu.de`

Abstract

Deductive reasoning plays a pivotal role in the formulation of sound and cohesive arguments. It allows individuals to draw conclusions that logically follow, given the truth value of the information provided. Recent progress in the domain of large language models (LLMs) has showcased their capability in executing deductive reasoning tasks. Nonetheless, a significant portion of research primarily assesses the accuracy of LLMs in solving such tasks, often overlooking a deeper analysis of their reasoning behavior. In this study, we draw upon principles from cognitive psychology to examine inferential strategies employed by LLMs, through a detailed evaluation of their responses to propositional logic problems. Our findings indicate that LLMs display reasoning patterns akin to those observed in humans, including strategies like *supposition following* or *chain construction*. Moreover, our research demonstrates that the architecture and scale of the model significantly affect its preferred method of reasoning, with more advanced models tending to adopt strategies more frequently than less sophisticated ones. Importantly, we assert that a model's accuracy, that is the correctness of its final conclusion, does not necessarily reflect the validity of its reasoning process. This distinction underscores the necessity for more nuanced evaluation procedures in the field.

Task Prompt:
Evaluate whether the given conclusion logically follows from the premises below:
1. There is a pink marble if and only if there is an olive marble.
2. If there is an olive marble then there is a brown marble.
3. Either there is a brown marble or else there is a black marble, but not both.

Conclusion: Either there is a pink marble or else there is a black marble, but not both.

Answer:
Okay, let's say there is a pink marble.
According to the first statement, there must be an olive marble. Then, according to the second statement, there must be a brown marble. But wait, that means there cannot be a black marble, because of the third statement!
So, the conclusion is actually true.

Answer:
1. If there is a pink marble, then there must be an olive marble.
2. If there is an olive marble, then there must be a brown marble.
3. If there is a brown marble, then there cannot be a black marble.

Therefore, if there is a pink marble, then there cannot be a black marble.
Conclusion: True

Figure 1: Given the propositional reasoning prompt (top box), the LLM shows two different inferential strategies: **supposition following** (left) and **chain construction** (right), see Section 2 for strategy details. Note that both answers are only partially correct, as the exclusive disjunction has only been proven for one of the cases (pink and not black). Model responses are generated by `LLaMA-2-Chat-70B` across two random seeds.

1 Introduction

Deductive reasoning, that is the process of drawing conclusions that logically follow from the information at hand, is an integral aspect of human cognition and plays a pivotal role in formulating sound and coherent arguments (Leighton, 2003). Take, for example, the following statements:

If there is a blue marble in the box then there is a green marble in the box.
There is a blue marble in the box.

Even without proper training in logic, most individuals can naturally deduce the valid conclusion:

Therefore, there is a green marble in the box.

This innate capability of drawing conclusions that invariably follow from the truth value of available information has been a focal point of scholarly interest for centuries (Holyoak and Morrison, 2005). Propositional logic, a subfield of deductive reasoning, focuses on constructing logical arguments based on the relationship between statements similar to those in the example previously mentioned (Hurley, 2011). Extensive research has been ded-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9370–9402

icated to examining human reasoning behavior in contexts that involve propositional logic. For instance, Van der Henst et al. (2002) have identified *five different strategies* people commonly employ when navigating problems of propositional logic (see Section 2). Such behavioral studies have been crucial in shaping theories that shed light on the fundamental elements of cognitive reasoning processes (Rips, 1994; Johnson-Laird, 1986; Kahneman et al., 1982).

In parallel, recent advancements in the field of large language models have demonstrated their potential in executing tasks involving deductive reasoning (Yang et al., 2023; Yu et al., 2024; Huang and Chang, 2023). Yet, the extent to which LLMs truly possess such abilities remains a subject of ongoing debate (Mahowald et al., 2024; Mitchell and Krakauer, 2023). Unlike behavioral studies in human reasoning that are often characterized by in-depth examinations of the reasoners' expressions, many studies on LLM-based reasoning tend to focus on task performance and accuracy metrics, offering limited insights into the underlying reasoning behavior of the models (Mitra et al., 2023; OpenAI et al., 2023; Team et al., 2023).

In this paper, we draw from the cognitive science literature (Van der Henst et al., 2002) and study inferential strategies employed by LLMs when solving propositional logic problems (see Figure 1). We analyze the reasoning behavior of three different language model families, varying in model size and fine-tuning procedure, and compare them to the behavior found in humans. To the best of our knowledge, we are the first to comprehensively compare inferential strategies employed by large language models and humans. We analyze the models' output both quantitatively and qualitatively via manual inspection, to provide insights into the soundness of their verbalized reasoning strategies. Our findings reveal that:

- All models exhibit inferential strategies akin to those observed in human reasoning, such as *supposition following* and *chain construction*.
- The inferential strategy employed is significantly influenced by the model family, as different families favor different approaches.
- Models are often right but for the wrong reasons: the *accuracy* of a model, that is the number of correct final conclusions, does not reflect whether its reasoning is *sound*, i.e. logically follows from the statements at hand.
- The strategy employed by a model is closely related to the *soundness* of its reasoning, where certain strategies lead to correct reasoning and others tend to introduce errors.
- In contrast to humans, models occasionally adopt a *symbolic strategy*, where formal logical calculus is employed to solve the propositional logic problem at hand.

Through this work, we hope to advance the understanding of reasoning in LLMs.

2 Strategies in Propositional Reasoning

Propositional logic studies the relationships among statements (or propositions) and the methods for constructing logical arguments based on them (Hurley, 2011). At the core of propositional logic are simple statements that can be combined through the use of logical connectives such as "not", "and", "or", and "if... then...", thereby forming more complex compound statements. Conclusions are logically deduced, where the truth value of the propositions necessitates the truth of the conclusion. This form of logical reasoning allows us to construct sound arguments that are invariably true, given the truth value of the information provided. As such, propositional logic is fundamental to various disciplines, including science, mathematics, and philosophy, where it offers a structured approach to reasoning and argumentation.

To gain insights into the inferential processes humans employ in propositional reasoning, Van der Henst et al. (2002) conducted a series of experiments that study the behavior of participants during propositional reasoning. They formulated straightforward propositional logic problems with neutral content (the presence or absence of colored marbles in a box, similar to the problem illustrated in Figure 1) and requested participants to articulate their thought processes while engaging with these problems. Participants were permitted the use of paper and pencil for their workings. Both their verbal explanations and written responses were meticulously recorded, transcribed and analyzed thereafter. Van der Henst et al. (2002) discovered five strategies reasoners commonly utilize to navigate the problems, offering insights into their inferential mechanisms employed during propositional reasoning. In the following, we give a short description of each strategy (illustrated in Figure 2). For more details and additional examples, we refer to the original study by Van der Henst et al. (2002).

Problem:
Evaluate whether the given conclusion logically follows from the premises below:
1. There is a blue marble if and only if there is a white marble.
2. Either there is a white marble or else there is a red marble, but not both.
3. There is a red marble if and only if there is a pink marble.
Conclusion: If there is a blue marble then there is a pink marble.

Incremental Diagram:
Blue iff white:
White
↕
Blue

White xor red:
White | Red
↕
Blue

Red iff pink:
White | Red ↔ Pink
↕
Blue

If blue then not pink.
Conclusion: False

Supposition Following:
Assuming we have a blue marble.
Then there is a white marble.
This means there is no red marble.
Thus there can not be a pink marble.

If blue then not pink.
Conclusion: False

Chain Construction:
If blue then white.
If white then not red.
If not red then not pink.

Therefore, if blue then not pink.
Conclusion: False

Compound Strategy:
1. Blue iff white.
2. White xor red.
- - - - - -
If blue then not red.
3. Red iff pink.
- - - - - -
If blue then not red and not pink.
Conclusion: False

Concatenation Strategy:
Blue iff white.
White xor red.
Red iff pink.

Blue iff (white xor (red iff pink)).

Therefore, blue xor pink.
Conclusion: False

Symbolic Strategy:
Blue iff white.
White xor red.
Red iff pink.

First, let's write down the statements in a logical form:
1. $B \leftrightarrow W$
2. $W \oplus R$
3. $R \leftrightarrow P$

Now, let's derive the conclusion using these statements [...].

Figure 2: An example for each of the five inferential strategies identified by Van der Henst et al. (2002) (to the left of the dashed vertical line) that human reasoners employ when solving tasks of propositional logic. Each strategy is illustrated by a single example adopted from the transcribed recordings published by the original study. In addition, we provide an example of the ● *symbolic strategy* occasionally encountered in LLMs (to the right of the dashed line). "Iff" denotes a biconditional, while "xor" indicates an exclusive disjunction.

Incremental Diagram. This strategy involves the creation of a comprehensive diagram that keeps track of all potential outcomes compatible with the premises of the problem. During the reasoning process, individuals progressively *increment* their diagrams to incorporate new information derived (see left box in Figure 2). The result is a single diagram that records a variety of possibilities compatible with the premises, often including even those that might be irrelevant to the task.[1]

Supposition Following. Reasoners employing this strategy start with a supposition, e.g. by assuming a marble of a certain color. Subsequently, they trace the implications of that supposition, logically following from the premises at hand, as illustrated in the upper second box from the left of Figure 2. The result is a sequence of literals (in this case, marbles of a certain color) without logical connectives.

The efficiency and success of *supposition following* strongly depends on the supposition made by the reasoner. While some suppositions lead to inferences that are relevant to the problem, others might lead to irrelevant conclusions.

Chain Construction. When employing this strategy, reasoners construct a chain of conditional statements derived either from the premises in the problem description or from intermediate deductions. An example of *chain construction* is displayed in the lower second box from the left of Figure 2. Premises are converted into a chain of conditional statements that are linked by their entities. A distinctive feature of this *chain* is the interconnection between conditionals, where the consequent of one conditional is the antecedent of the following.

Compound Strategy. Reasoners following the *compound strategy* combine two or more statements to derive a new *compound conclusion*. This process yields a series of novel conclusions, each building upon the preceding ones. An illustrative example of this strategy is given in the upper second box from the right of Figure 2. Based on the first two premises, the compound conclusion: *"If blue then not red."* is inferred, and then used

[1]In contrast to Van der Henst et al. (2002), we in fact observe no single occurrence of the *incremental diagram strategy* in LLMs, despite the authors finding that this strategy is most frequently employed by humans. We believe that this discrepancy stems from the use of pen and paper in human assessments, implicitly encouraging diagrammatic reasoning. Exploring how this observation changes with vision-language models would be an intriguing area for future research.

to draw another compound conclusion (*"If blue then not red and not pink."*) together with the last premise of the problem statement.

Concatenation Strategy. This approach entails the concatenation of two or more statements into a *single conclusion* encompassing the logical implications of each combined proposition. This strategy is subtle and has only been infrequently observed by Van der Henst et al. (2002). An example of the strategy is illustrated in the lower second box from the right of Figure 2.

Symbolic Strategy. We could identify an additional strategy occasionally employed by LLMs, which has not been observed by Van der Henst et al. (2002) in human reasoners. This strategy, which we denote as *symbolic strategy*, is characterized by models employing formal logical calculus to solve the tasks at hand. When following this strategy, models either translate logical statements that are expressed in natural language (e.g. *"If there is a white marble then there is not a red marble."*) into formal logic ($W \rightarrow \neg R$), and then operate on those expressions, or create a truth table from which they aim to infer the validity of the conclusion. An illustration of this strategy is provided in the right box of Figure 2.

3 Experimental Setup

Task Overview. Our task setup aligns with the experiment conducted by Van der Henst et al. (2002) to allow for a fair comparison between the inferential strategies found in humans and those identified in LLMs.[2] In particular, we evaluate each model on the 12 problems of propositional logic suggested by Van der Henst et al. (2002) (an overview of each problem can be found in Figure 5 in the appendix). For each problem, models are presented with a set of statements (or premises) and must determine whether a given conclusion logically follows (for an example, see Figure 1). Eight out of 12 problems involve three premises and a conclusion, while the remaining four problems consist of four premises leading to a conclusion. All premises, as well as the conclusions resemble either biconditionals, exclusive disjunctions or conditionals. Two problems (4 and 6) include a redundant first premise. All premises are stated such that two subsequent statements contain one proposition in common, except of two problems (11 and 12), which are arranged in a non-sequential manner. For half of the problems, the conclusions logically follow from the premises, whereas for the other half, they do not. To avoid the influence of external knowledge and ensure content neutrality, Van der Henst et al. (2002) framed the problems around the presence of colored marbles in a box, with colors assigned randomly to each entity within a problem.

Language Models. We aim to investigate various factors that might impact the inferential strategies displayed by LLMs. These factors include the type of model, its size, and the emphasis on alignment during training (Tunstall et al., 2023). Therefore, we assess a total of five models, consisting of three prominent open-access model types: Llama 2 (Touvron et al., 2023) with model sizes of 7B, 13B, and 70B, the recently released Mistral-7B model (Jiang et al., 2023), and Zephyr-7B (Tunstall et al., 2023), an extension of Mistral-7B with a focus on intent alignment through fine-tuning with AI Feedback (AIF). For our evaluations, we utilize the publicly accessible model weights from the HuggingFace platform, specifically `Llama-2-chat-hf`[3](7B, 13B, and 70B), `Mistral-7B-Instruct-v0.2`,[4] and `zephyr-7b-beta`.[5] We consciously opt not to include proprietary models accessible via paid APIs, despite their reported superior performance in reasoning tasks (Team et al., 2023). This methodological choice reflects our commitment to promoting transparent and reproducible scientific research. Note that in this work, we refer to the above models when using abbreviations such as LLaMA-2, Mistral-7B-Instruct or Zephyr-7B-β.

Evaluation Setup. We prompt each model with a system message providing context about the task they are about to solve and the format in which they should answer (for the full prompt, see Figure 5 in the appendix). Analogous to Van der Henst et al. (2002), we inform the model of its participation in an experiment designed to explore reasoning processes, and instruct it to *"think aloud"* as it tackles the problem. In addition to the system message, we provide a user prompt that contains the problem description. In cases where the model does not accept system messages (such as `Mistral-7B-Instruct-v0.2`), we prepend the content of the system message to the user prompt. To prevent biasing the model towards a certain strat-

[2] More specifically, experiment one of Van der Henst et al. (2002).

[3] https://huggingface.co/meta-llama

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct

[5] https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

Problem 1:

Statements:
1. Either there is a white marble in the box or else there is a black marble in the box, but not both.
2. Either there is a black marble in the box or else there is a pink marble in the box, but not both.
3. There is a pink marble in the box if and only if there is a gray marble in the box.

Conclusion: If there is a white marble in the box then there is a gray marble in the box.

Response:	**Annotator Comments**:
Sure, I'd be happy to help! Here's my reasoning process for this problem:	
1. First, I'll start by analyzing the given statements:	
From Statement 1, we know that the box contains either a white marble or a black marble, but not both.	Reformulation of statement 1.
From Statement 2, we know that the box contains either a black marble or a pink marble, but not both.	Reformulation of statement 2.
From Statement 3, we know that if there is a pink marble in the box, then there must be a gray marble in the box.	Reformulation of statement 3.
2. Now, let's consider the conclusion: If there is a white marble in the box then there is a gray marble in the box.	Repetition of conclusion.
Can we logically deduce this conclusion from the given statements?	Chain construction (sound).
Well, if there is a white marble in the box, then by Statement 1, there cannot be a black marble in the box.	W → ¬ B
If there is no black marble in the box, then by Statement 2, there must be a pink marble in the box.	¬ B → P
But if there is a pink marble in the box, then by Statement 3, there must be a gray marble in the box.	P → G
So, yes, we can logically deduce the conclusion that if there is a white marble in the box, then there is a gray marble in the box.	Conclusion: W → G
Conclusion: True	Final answer (correct).

Figure 3: The response (lower left box) of LLaMA-2-70B to problem 1 (top box) of the problem set, demonstrating **chain construction**. The model correctly constructs a chain of conditionals (highlighted in yellow within the model's response) based on the premises, leading from the antecedent of the final conclusion to its consequent. Comments made by the annotators are presented in the adjacent right panel.

egy, we refrain from providing few-shot examples, as done also by Leidinger et al. (2023). Instead, we elicit reasoning through zero-shot chain-of-thought prompting (*"Let's think step by step"*) (Kojima et al., 2022). Answers for each model are generated with nucleus sampling using `Llama-2-chat-hf`'s default values (top-$p = 0.9$, temperature $T = 0.6$), as we found this configuration to work well for all models. To account for the statistical nature of language models, we ask each model to solve the set of propositional problems across 5 random seeds, resulting in a total of 60 responses per model. Our code is publicly available at: https://github.com/mainlp/inferential-strategies.

We record all answers and manually evaluate them (a total of 300 responses) for strategies employed in their reasoning (see Figure 3 for an example). For each model response, we qualitatively evaluate for *strategy* and *soundness*. That is, we manually label the inferential strategies identified, and the logical validity of the model's reasoning. In addition, we record whether the final answer is correct. In cases of faulty reasoning, we categorize the type of error. This comprehensive manual evaluation of model responses is independently conducted by two hired students with expertise in manual data annotation. To gauge the quality of the annotations, we report an overall Cohen's Kappa value of $\kappa = 0.98$. For details on the inter-annotator agreement of each label, we refer to Table 2 in the appendix. Further annotated examples can be found in Appendix C. Following the recommendations put forward by Leidinger et al. (2023), we make all input prompts, model responses and manual annotations publicly available at: huggingface.co/datasets/mainlp/inferential_strategies.

4 Results and Analysis

In this section, we present the results of our evaluation. We begin with a quantitative analysis of the inferential strategies employed by LLMs, as well as the logical validity of their reasoning. This is followed by a qualitative analysis providing a more in-depth examination of the models' reasoning.

Model	Supposition Following	Chain Construction	Compound Conclusion	Concatenation Strategy	Symbolic Strategy	Correct Answer	Sound Reasoning
Zephyr-7B-β	**60.0%** (*55.1*)	18.3% (*17.3*)	10.0% (*8.9*)	1.7% (*1.4*)	20.0% (*17.3*)	45.0 ± 15.5	25.0 ± 10.5
Mistral-7B-Instruct	**35.0%** (*38.4*)	10.0% (*10.7*)	**35.0%** (*38.4*)	3.3% (*3.4*)	8.3% (*9.1*)	55.0 ± 10.0	25.0 ± 7.5
LLaMA-2-7B	**20.0%** (*50.2*)	**20.0%** (*30.2*)	6.7% (*10.9*)	3.3% (*5.4*)	1.7% (*3.3*)	46.7 ± 6.7	0.0 ± 0.0
LLaMA-2-13B	28.3% (*35.7*)	**36.7%** (*46.9*)	6.7% (*8.7*)	6.7% (*8.7*)	0.0% (*0.0*)	40.0 ± 8.2	15.0 ± 6.2
LLaMA-2-70B	45.0% (*42.3*)	**50.0%** (*46.8*)	3.3% (*2.9*)	1.7% (*1.8*)	6.7% (*6.2*)	56.7 ± 6.2	31.7 ± 9.7
Human Reasoner†	– (21.0)	– (25.0)	– (19.0)	– (0.0)	– (0.0)	100 ± 0.0	–

Table 1: Relative occurrences of inferential strategies employed by the different language models when solving the problems of propositional logic. All values reflect average percentages, calculated over five random seeds, with standard deviations reported in Table 4 in the appendix. Strategies that a model favors are highlighted in bold. Values in parentheses denote fractions with respect to the total number of strategies employed by that model. Values of correct answers and instances of sound reasoning are reported with their standard deviations. †The comparison with human reasoners is based on findings by Van der Henst et al. (2002), where dashes denote missing values.

4.1 Quantitative Analysis

Table 1 provides an overview of the frequencies with which large language models employ inferential strategies when navigating the problems of propositional logic described in Section 3. Our evaluation reveals that all models display strategies akin to those observed by Van der Henst et al. (2002). In particular, we find that, similar to humans, models commonly employ *supposition following*, *chain construction* and the *compound strategy*. In addition, we observe that models occasionally utilize the *symbolic strategy*, employing techniques from logical calculus to solve the tasks (see Section 2). Note that, similar to humans, models might switch from one strategy to another during a single problem, demonstrating multiple strategies within their responses (see Figure 19 in the appendix for an example). Surprisingly, we observe that distinct model families favor different inferential strategies. For instance, Zephyr-7B-β predominantly employs *supposition following*, while Mistral-7B-Instruct is equally inclined towards drawing *compound conclusions*. In contrast, models from the Llama 2 series tend to rely on *supposition following* and *chain construction*, with negligible use of the *compound strategy*. Our analysis further reveals a discrepancy between the correctness of the models' final answers and the logical soundness of their reasoning. While all models achieve an answer accuracy that approximately coincides with chance in our experimental setup, an analysis of their reasoning validity reveals a different picture: LLaMA-2-70B outperforms the other models by reasoning correctly in about 31.7% of cases, while Zephyr-7B-β and Mistral-7B-Instruct produce sound reasoning in 25% of the problems. We note that all models perform rather poorly on the propositional tasks, with LLaMA-2-7B failing entirely to construct sound arguments.

Human Reasoning. Van der Henst et al. (2002) compute the percentages with which human reasoners employ inferential strategies with respect to the total number of strategies observed in their experiment, and not with respect to the total number of problems considered. Thus, their reported values mainly reflect which strategies are favored more or less by the reasoner, but do not provide information about how frequently a strategy has been observed in the overall context. To make our findings comparable to the results of Van der Henst et al. (2002), we convert our results respectively (see values in parentheses in Table 1). We note that almost all models seem to favor *supposition following* to a higher degree than human reasoners, who employ this strategy in only about 21% of overall use. In contrast, humans seem to draw compound conclusions more readily, except for Mistral-7B-Instruct which shows a tendency more than twice as high. Overall, both LLMs and humans hardly employ the *concatenation strategy*. Interestingly, Van der Henst et al. (2002) report that all reasoners successfully solve the problems of propositional logic, though not always for the correct reasons. While the study does not provide data on the number of problems where humans reasoned correctly, the high success rate of human participants contrasts sharply with the performance of the models.

Effect of Model Size. Our evaluation of the Llama 2 series across three different model sizes—7B, 13B, and 70B parameters—demonstrates that

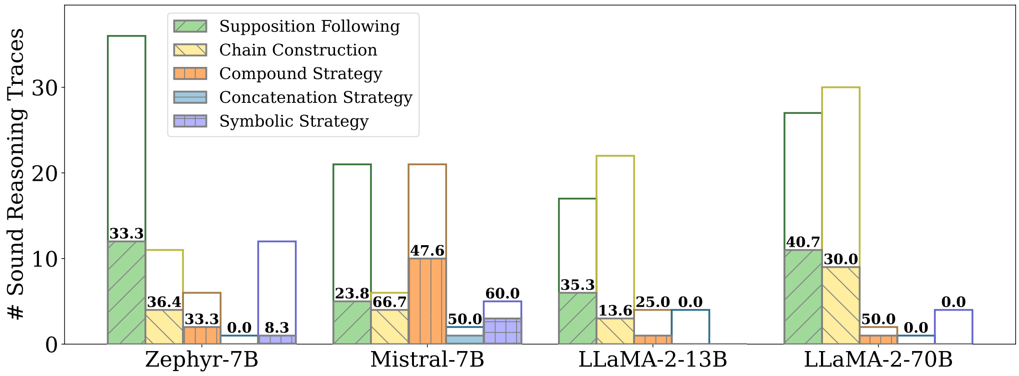

Figure 4: Instances where models generate sound reasoning traces that logically follow from the problem statement. For each inferential strategy, the ratio of sound reasoning traces (represented by the filled portion) to the overall application of that strategy (denoted by the unfilled bar) is depicted. Ratios are expressed as percentages above the corresponding filled section. Note that LLaMA-2-7B is not displayed as it does not exhibit sound reasoning.

model scale significantly influences the frequency with which strategies are employed by the model. In particular, we observe that with increasing model size, Llama 2 employs strategies more readily. Furthermore, larger models within the Llama 2 framework are observed to generate a greater number of sound reasoning traces. We interpret this trend as a result of the model's improving proficiency in strategic reasoning as its scale increases.

Effect of Alignment. The alignment of a model's response with human preferences is crucial to emulate human-like behavior (Ouyang et al., 2022). Zephyr-7B-β is an iteration of Mistral-7B that is fine-tuned with AI Feedback (AIF) for improved intent alignment (Tunstall et al., 2023). In comparison to the observations made by Van der Henst et al. (2002), where besides the *incremental diagram strategy* (34%), *chain construction* was employed most frequently by humans, Zephyr-7B-β demonstrates a marked preference for *supposition following* and significantly less engagement in *chain construction*. Moreover, it is noteworthy that among the evaluated models, Zephyr-7B-β most frequently adopts the *symbolic strategy*, an approach not reported in human reasoners.

Sound Reasoning. As previously highlighted, the accuracy of a model's final answer does not necessarily serve as a reliable indicator of its reasoning capability. In particular, we observe that models often arrive at correct answers, but through flawed reasoning processes (refer to Figure 10 in the appendix for an illustration). Interestingly, we also find instances where models provide incorrect final answers despite reasoning correctly (for an example, see Figure 16 in the appendix). Our analysis reveals only a moderate positive correlation between the accuracy of the models' final answers and the logical soundness of their reasoning, with a Pearson correlation coefficient $r(298) = 0.45$ and a statistically significant p-value of less than $0.0001 (p = 1.6 \times 10^{-16})$. This observation aligns with findings from previous studies (Ye and Durrett, 2022; Creswell and Shanahan, 2022) and underscores the need for more nuanced evaluation procedures, particularly in multiple-choice settings, where models might select the correct answer by chance rather than through rigorous reasoning.

In Figure 4, we explore the relationship between the inferential strategies employed by the models and the validity of their reasoning. For each strategy, we quantify the proportion of instances where the models' reasoning is sound, compared to the overall application of that strategy. Our analysis reveals variability in the effectiveness with which different models apply various strategies. For example, Mistral-7B-Instruct tends to reason correctly when using approaches such as the *chain*, *compound*, or *symbolic strategy*, yet frequently encounters reasoning errors with *supposition following*. On the other hand, LLaMA-2-70B exhibits proficiency in *supposition following*, but struggles with the *symbolic strategy*.

4.2 Qualitative Analysis

We supplement our quantitative analysis by a more detailed qualitative analysis of the models' reasoning behavior. Figure 3 depicts LLaMA-2-70B's response to problem 1 of the task set. The response illustrates a frequently observed behavior. Initially, models tend to analyze the problem's propositions, often by paraphrasing each premise and the conclusion to be evaluated. They then embark on a reasoning process, typically utilizing one of the previously mentioned strategies. In the example, LLaMA-2-70B employs *chain construction*, creating a logical chain of conditionals that leads from the antecedent of the final conclusion to its consequent, thereby correctly affirming the conclusion's logical validity. A notable pitfall in such reasoning chains is the models' occasional misinterpretation of logical negations, leading to erroneous chains like: A $\rightarrow \neg$ B; B $\rightarrow$ C; therefore A $\rightarrow$ C, where the negation in the first conditional is overlooked (for an illustrative case, refer to Figure 11 in the appendix). This behavior can be found across all models and aligns with previous work reporting difficulties of LLMs in understanding logical negations (Truong et al., 2023).

When employing *supposition following*, models often fail to consider all implications of their assumptions. Instead, they tend to focus only on immediate inferences, while overlooking further consequences crucial for assessing the conclusion's validity. This leads to models prematurely concluding the inability to definitively determine the logical validity of the final conclusion: *"Based on our analysis, we cannot definitively say that the conclusion logically follows from the given statements"* (see Figure 7 in the appendix for a respective example). Another source of error in *supposition following* involves models making improper suppositions, such as conjecturing about a marble not mentioned in the final conclusion, and deriving disjointed intermediate conclusions that do not aid in solving the problem. An example of this behavior can be found in Figure 8 in the appendix.

Finally, we identify two behaviors in models that mirror logical errors seen in human reasoners (Van der Henst et al., 2002). First, models frequently attempt to prove an exclusive disjunction (A $\oplus$ B) by only considering a single conditional case (A $\rightarrow \neg$ B), and second, they sometimes engage in the logical fallacy known as denial of the antecedent: A $\rightarrow$ B; therefore $\neg$ A $\rightarrow \neg$ B (for illustrative examples, see Figures 12 and 13 in the appendix, respectively).

5 Related Work

Human Strategies in Deductive Reasoning. A considerable amount of research, especially within psychology and cognitive science, has explored how humans approach deductive reasoning tasks (Schaeken et al., 2000). A prominent focus of these studies is on heuristics, which are cognitive shortcuts that individuals employ to arrive at satisfactory conclusions in deductive reasoning despite potential flaws in the underlying logic (Kahneman et al., 1982; Evans, 1989; Gigerenzer and Todd, 1999; Davis, 2018). For instance, Woodworth and Sells (1935) demonstrate that individuals tend to accept conclusions in syllogistic reasoning as valid when they share logical quantifiers with the premises, regardless of their actual logical validity. Nonetheless, such reliance on heuristics can result in errors and falls short of the level of strategic reasoning necessary to develop sound and coherent arguments (Kahneman, 2012). Further research has delved into more sophisticated strategies utilized by individuals in deductive reasoning. Based on the mental model theory (Johnson-Laird, 1986), Bucciarelli and Johnson-Laird (1999) identify a variety of strategies commonly employed by individuals in syllogistic reasoning. Byrne and Handley (1997) study strategies of individuals in *knight-and-knave* puzzles, where the truthfulness of statements made by hypothetical characters have to derived. Their experiments reveal that humans engage in both forward and backward inferences to navigate through potential solutions.

Human Reasoning Behavior in LLMs. Recent research has started to explore the extent to which LLMs mirror human-like reasoning behaviors. Dasgupta et al. (2023) demonstrate content-effects akin to those observed in human reasoning, where the deductive process is influenced by the content of the problem statement. Eisape et al. (2023) find that LLMs, similar to humans, exhibit biases such as ordering effects in syllogistic reasoning tasks. Several other studies have delved into the prevalence of biases and heuristics within LLMs (Binz and Schulz, 2023; Talboy and Fuller, 2023; Shaki et al., 2023; Suri et al., 2024). However, to the best of our knowledge, we are the first who study the presence of more sophisticated human strategies in the context of LLM-based deductive reasoning.

Faithful Reasoning with LLMs. Large language models can be instructed to explain the reasoning process by which they derive their final conclusions (Wei et al., 2022; Kojima et al., 2022). However, several studies indicate that these self-explanations might not always be *faithful*, i.e. accurately represent the model's underlying reasoning process

(Jacovi and Goldberg, 2020; Agarwal et al., 2024; Lyu et al., 2024). For instance, Turpin et al. (2023) demonstrate that LLMs such as GPT-3.5 (OpenAI, 2023) and Claude 1 (Anthropic, 2023) often fail to mention biasing features in their input that significantly influence their decisions. Instead, the models produce *plausible* yet misleading explanations that give a false account of the underlying decision process. Lanham et al. (2023) probe the faithfulness of explanations by evaluating how the final conclusions of LLMs change when rationales are truncated or errors are introduced. Their findings reveal that the extent to which models rely on their rationales varies strongly across models and tasks. Matton et al. (2024) propose a method to quantify the faithfulness of explanations based on high-level concepts in the models' input that influence decision-making. By measuring the difference between the set of concepts that LLMs deem influential and the set that truly are, instances of unfaithfulness could be identified, including cases where LLMs overlook the impact of social biases in their decision-making processes. Another important consideration is whether the model's final conclusion aligns with its preceding explanation. As highlighted in Section 4.1, a correct conclusion might not always be the product of a logically sound reasoning trace, particularly in multiple-choice setups. Conversely, a sound rationale may not always lead to a logically consistent answer. Related work by Ye and Durrett (2022) indicates that in question-answering and natural language inference tasks, explanations generated by LLMs such as OPT (Zhang et al., 2022) and GPT-3 (Brown et al., 2020) often do not entail the models' final conclusions. Further studies aim to enhance the models' faithfulness, for instance by enforcing causality from proof generation to entailment prediction. This can be achieved by either restricting the model's context (Sanyal et al., 2022; Creswell and Shanahan, 2022; Radhakrishnan et al., 2023), or by utilizing deterministic tools that are inherently faithful by design (Lyu et al., 2023).

6 Conclusion

In this paper, we examine the inferential strategies employed by LLMs in solving problems of propositional logic. Through a comprehensive evaluation of their reasoning behavior, we demonstrate that LLMs adopt strategies akin to those observed in human reasoners. Our quantitative analysis reveals that the frequency with which a model adopts a specific strategy strongly depends on its type, size, and fine-tuning procedure. Moreover, our analysis suggests that the accuracy of a model's final conclusions does not adequately capture its reasoning capabilities, underscoring the importance of a more sophisticated evaluation framework that includes the model's reasoning paths. We also provide a qualitative analysis of typical reasoning behaviors among models, pinpointing prevalent errors such as difficulties in understanding negations or recognizing all implications of a supposition.

7 Limitations

While our work contributes to the understanding of reasoning processes in large language models by demonstrating that these models employ inferential strategies in propositional logic similar to humans, it encompasses several limitations that could be addressed in future work.

Task setup. Our study is constrained by a limited set of problems, designed within a fixed framework that revolves around hypothesis validation based on 3-4 statements of propositional logic. We employ a constant and neutral content, disregarding potential content-effects on the models' reasoning behavior, as shown by Dasgupta et al. (2023). Similarly, we have not yet examined factors such as the complexity of the problems, the differences between hypothesis validation and generation, and the impact of logical connectives utilized in the premises. We believe that these factors are worth investigating and leave a detailed examination to future work.

Evaluation Framework. The extent of our manual evaluation is limited by both the number of samples reviewed and the quantity of annotators involved. Despite our efforts to maximize the use of available resources, these constraints may affect the scalability and reliability of our results. Additionally, we instruct all models through zero-shot chain-of-thought prompting (*"Let's think step by step"*) (Kojima et al., 2022). Exploring alternative reasoning frameworks, such as Tree of Thoughts (Yao et al., 2023) or Graph of Thoughts (Besta et al., 2024), could provide valuable insights into their influence on model behavior and the inferential strategies adopted. Based on our annotated data, we endeavored to develop a classifier capable of automatically identifying the inferential strategies employed in the models' output, which was intended to complement our manual evaluation setup. However, due to the complexity of the task and limited size of our annotated dataset, our classifier struggled with generalization to new, unseen responses. In future endeavors, we aim to allocate more resources towards expanding our manual annotation efforts and explore this direction further. Finally, our study predominantly offers a behavioral analysis and does not delve into the mechanistic aspects that might explain the diversity in strategy usage by the models. Investigating how model-internal mechanisms might influence their choice of reasoning strategy presents a compelling direction for future research.

Acknowledgements

We would like to thank the MaiNLP lab members for their insightful feedback, specifically Diego Frassinelli, Rob van der Goot, Siyao Peng, Robert Litschko, Jian Lan, Daniela Teodorescu, Xinpeng Wang, Verena Blaschke, Elena Senger, Elisa Bassignana, and Max Müller-Eberstein. Furthermore, we would like to express our gratitude to Huangyan Shan for her valuable work and support in data annotation. Our appreciation extends to the anonymous reviewers for their comments and suggestions. Lastly, we acknowledge the support for BP through the ERC Consolidator Grant DIALECT 101043235.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. ArXiv:2402.04614 [cs].

Anthropic. 2023. Introducing claude. *Anthropic Research*. Available at https://www.anthropic.com/news/introducing-claude. Accessed: 2024-05-26.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120. ArXiv:2206.14576 [cs].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Monica Bucciarelli and P.n. Johnson-Laird. 1999. Strategies in Syllogistic Reasoning. *Cognitive Science*, 23(3):247–303.

R. M. Byrne and S. J. Handley. 1997. Reasoning strategies for suppositional deductions. *Cognition*, 62(1):1–49.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. ArXiv:2208.14271 [cs].

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2023. Language models show human-like content effects on reasoning tasks. ArXiv:2207.07051 [cs].

Andrew M. Davis. 2018. Biases in Individual Decision-Making. In *The Handbook of Behavioral Operations*, pages 149–198. John Wiley & Sons, Ltd.

Tiwalayo Eisape, M. H. Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van Steenkiste, and Tal Linzen. 2023. A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models. ArXiv:2311.00445 [cs].

Jonathan St. B. T. Evans. 1989. *Bias in human reasoning: Causes and consequences*. Bias in human reasoning: Causes and consequences. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US. Pages: ix, 145.

Gerd Gigerenzer and Peter M. Todd. 1999. *Simple heuristics that make us smart*. Simple heuristics that make us smart. Oxford University Press, New York, NY, US. Pages: xv, 416.

Keith J. Holyoak and Robert G. Morrison. 2005. *The Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press. Google-Books-ID: zn-bkHaC8QeMC.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Patrick J. Hurley. 2011. *A Concise Introduction to Logic*, 11th edition edition. CENGAGE Learning Custom Publishing, Boston, MA.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

P. N. Johnson-Laird. 1986. *Mental models: towards a cognitive science of language, inference, and consciousness*. Harvard University Press, USA.

Daniel Kahneman. 2012. *Thinking, Fast and Slow: Daniel Kahneman*, 1st edition edition. Penguin, London.

Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. ArXiv:2307.13702 [cs].

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Jacqueline P. Leighton. 2003. Defining and Describing Reason. In Jacqueline P. Leighton and Robert J. Sternberg, editors, *The Nature of Reasoning*, pages 3–11. Cambridge University Press, Cambridge.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics*, pages 1–67.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Katie Matton, Robert Ness, and Emre Kiciman. 2024. Walk the talk? measuring the faithfulness of large language model explanations. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching Small Language Models How to Reason. ArXiv:2311.11045 [cs].

OpenAI. 2023. Models overview. *OpenAI Research*. Available at `https://platform.openai.com/docs/models`. Accessed: 2024-05-26.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Question decomposition improves the faithfulness of model-generated reasoning. ArXiv:2307.11768 [cs].

Lance J. Rips. 1994. *The Psychology of Proof: Deductive Reasoning in Human Thinking*. The MIT Press.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. FaiRR: Faithful and robust deductive reasoning over natural language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1075–1093, Dublin, Ireland. Association for Computational Linguistics.

Walter Schaeken, Gino De Vooght, André Vandierendonck, and Géry d'Ydewalle, editors. 2000. *Deductive reasoning and strategies*. Deductive reasoning and strategies. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Pages: xiv, 321.

Jonathan Shaki, Sarit Kraus, and Michael Wooldridge. 2023. Cognitive effects in large language models. In *ECAI 2023*, pages 2105–2112. IOS Press.

G. Suri, L. R. Slater, A. Ziaee, and M. Nguyen. 2024. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *Journal of Experimental Psychology: General*, 153(4):1066–1075.

Alaina N. Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption. ArXiv:2304.01358 [cs].

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha

Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuoyiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin

Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G. Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, T. J. Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate,

James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M. K. Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A Family of Highly Capable Multimodal Models. ArXiv:2312.11805 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. ArXiv:2310.16944 [cs].

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Jean-Baptiste Van der Henst, Yingrui Yang, and P.n. Johnson-Laird. 2002. Strategies in sentential reasoning. *Cognitive Science*, 26(4):425–468.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

R. S. Woodworth and S. B. Sells. 1935. An Atmosphere Effect in Formal Syllogistic Reasoning. *Journal of Experimental Psychology*, 18(4):451.

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023. Logical Reasoning over Natural Language as Knowledge Representation: A Survey. ArXiv:2303.12023 [cs].

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392. Curran Associates, Inc.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Comput. Surv.* Just Accepted.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. ArXiv:2205.01068 [cs].

A Additional Experimental Details

In this section, we provide additional details about the experimental setup, including supplementary information about the problem formulations and prompts utilized.

A.1 Task Prompts

Figure 5 displays the task prompt and problem formulations employed in assessing the language models described in Section 3. Note that the prompt

[INST] «SYS»
You participate in an experiment that tries to understand how people reason.
Your task is to solve logical reasoning problems. In particular, you are given set of statements and your task is to say whether a conclusion logically follows from the statements.
Please answer with 'True' or 'False' for each conclusion. In addition, it is important that you think-aloud as you tackle the problem and report every step in your reasoning process.
«/SYS»

Instruction
Explain whether the conclusion logically follows from the set of statements below. Please report all your reasoning steps.
End your reasoning with: Conclusion: True/False.

Statements:
<statements and conclusion from below>

Let's think step by step. [/INST]

Problem 1:
Statements:
1. White xor black.
2. Black xor pink.
3. Pink iff gray.

Conclusion: If white then gray.

Problem 2:
Statements:
1. Brown iff orange.
2. Orange xor yellow.
3. Yellow iff green.

Conclusion: If brown then green.

Problem 3:
Statements:
1. Green iff purple.
2. If purple then gray.
3. Gray xor yellow.

Conclusion: Green xor yellow.

Problem 4:
Statements:
1. Red xor maroon.
2. Maroon xor yellow.
3. Yellow iff orange.

Conclusion: If maroon then orange.

Problem 5:
Statements:
1. Purple iff yellow.
2. Yellow iff blue.
3. Blue xor orange.

Conclusion: Purple xor orange.

Problem 6:
Statements:
1. Gray iff yellow.
2. Yellow xor olive.
3. Olive iff black.

Conclusion: If yellow then black.

Problem 7:
Statements:
1. Blue iff red.
2. Red xor white.
3. White iff pink.

Conclusion: If not blue then pink.

Problem 8:
Statements:
1. Olive xor brown.
2. Brown iff gray.
3. Gray xor maroon.

Conclusion: If not olive then maroon.

Problem 9:
Statements:
1. Purple iff blue.
2. Blue iff olive.
3. Olive xor red.
4. Red xor green.

Conclusion: If purple then green.

Problem 10:
Statements:
1. Brown iff yellow.
2. Yellow xor green.
3. Green iff purple.
4. Purple iff olive.

Conclusion: If brown then olive.

Problem 11:
Statements:
1. Red iff maroon.
2. Green xor olive.
3. Maroon iff green.
4. Olive xor brown.

Conclusion: If red then brown.

Problem 12:
Statements:
1. Blue iff brown.
2. White iff green.
3. Brown xor white.
4. Green iff purple.

Conclusion: If blue then purple.

Figure 5: The task prompt (upper yellow box) as well as statements and conclusion for each propositional logic problem (lower gray boxes). In the task prompt, the placeholder *"<statements and conclusion from below>"* is replaced with the actual statements and conclusion relevant to each problem. To enhance readability, we employ abbreviations within the problem statements. In the actual prompt, "colorA iff colorB" is replaced by "There is a colorA marble in the box if and only if there is a colorB marble in the box". Similarly, "colorA xor colorB" denotes "Either there is a colorA marble in the box or else there is a colorB marble in the box, but not both". Lastly, "If colorA then colorB" stands for "If there is a colorA marble in the box then there is a colorB marble in the box".

	Supposition Following	Chain Construction	Compound Conclusion	Concatenation Strategy	Symbolic Strategy	Correct Answer	Sound Reasoning
Zephyr-7B-β	1.0	0.94	1.0	1.0	1.0	1.0	1.0
Mistral-7B-Instruct	1.0	0.9	1.0	1.0	1.0	1.0	1.0
LLaMA-2-7B	0.89	0.95	1.0	0.79	1.0	1.0	1.0
LLaMA-2-13B	0.88	1.0	0.85	1.0	1.0	1.0	1.0
LLaMA-2-70B	0.97	1.0	1.0	1.0	1.0	1.0	1.0

Table 2: Cohen's Kappa values to assess the inter-annotator agreement across different models and label categories.

template, i.e. special tokens and their arrangements, might vary depending on the specific language model used. Within the task prompt (provided in the upper box), the problem statements and conclusion for a given problem are replaced with the corresponding problem formulations found in the lower gray boxes. In the final version of the prompt, the phrase "colorA iff colorB" is expanded to "There is a colorA marble in the box if and only if there is a colorB marble in the box". Similarly, "colorA xor colorB" is interpreted as "There is either a colorA marble or a colorB marble in the box, but not both", and "If colorA then colorB" is articulated as "If there is a colorA marble in the box, then there is a colorB marble in the box".

A.2 Annotator Instructions

Our assessment of model responses involves a comprehensive independent review by two students who are specialized in the field of natural language processing and have expertise in manual data annotation. To ensure a high quality of annotations, we offer comprehensive training to both annotators. This training includes detailed explanations and extensive examples of the strategies identified by Van der Henst et al. (2002), complemented by a session dedicated to clarifying any questions that may emerge. Subsequently, the annotators are tasked with independently annotating practice examples, which serves to highlight and address any ambiguities in the annotation process. Only when both annotators are confident in their understanding of each strategy do we proceed. We instruct both annotators to independently go through each model response and mark parts where they identify a certain strategy to be employed. Each strategy is marked in a unique color code, which is afterwards converted into labels that signify the use of a particular strategy. In addition, we instruct both annotators to label whether the reasoning is sound, and the final conclusion of the model is correct. Furthermore, we ask them to classify any logical errors identified within the reasoning process. To maintain a high standard of annotation quality, annotators are instructed to review the model responses twice.

A.3 Inter-Annotator Agreement

To assess the reliability of our manual evaluation process (see Section 3), we quantify the inter-annotator agreement by calculating Cohen's Kappa for each category and model, as illustrated in Table 2. Generally, the results indicate an almost perfect level of agreement across all categories and models, with Cohen's Kappa values ranging from $0.81 \leq \kappa \leq 1.0$. An exception is observed in the case of the *concatenation strategy* applied by LLaMA-2-7B, for which we report a substantial agreement level, with a Kappa value of $\kappa = 0.79$, slightly below the threshold for almost perfect agreement.

A.4 Model Details

We report further details about the models used in this study in Table 3. In particular, we provide information about the number of parameters, context length and fine-tuning procedure for each model.

B Additional Quantitative Results

In this segment, we present supplementary findings from our quantitative evaluation. Table 4 illustrates the frequencies with which the different language models employ inferential strategies when navigating the problems of propositional logic, as outlined in Section 3. Values denote percentages averaged across five distinct random seeds, accompanied by their standard deviation. Furthermore, we detail the proportions of correct final conclusions and sound reasoning. Note that all percentages are calculated relative to the overall count of tasks within the experimental framework.

C Annotated Model Responses

Within this section, we showcase examples of model responses that exemplify each inferential

Model	Base Model	Parameters	Context Length	Tokens	Fine-tuning
Zephyr-7B-β	Mistral	7B	8192 tokens	-	dSFT, AIF
Mistral-7B-Instruct	Mistral	7B	8192 tokens	-	SFT
LLaMA-2-7B-Chat	LLaMA-2	7B	4K tokens	2.0T	SFT, RLHF
LLaMA-2-13B-Chat	LLaMA-2	13B	4K tokens	2.0T	SFT, RLHF
LLaMA-2-70B-Chat	LLaMA-2	70B	4K tokens	2.0T	SFT, RLHF

Table 3: Properties of the models used in this study. The context length refers to the base model's training. Tokens relate to the number of tokens in the pre-training data only. We use the following abbreviations for the fine-tuning procedure: supervised fine-tuning (SFT), reinforcement learning with human feedback (RLHF), distilled supervised fine-tuning (dSFT), and AI feedback through preferences (AIF). Information about the Llama 2 family is taken from Touvron et al. (2023), specifications for Mistral-7B-Instruct are provided by Jiang et al. (2023). For Zephyr-7B-β, we consider the work of Tunstall et al. (2023). Dashes represent cases in which we could not find the respective information.

strategy identified in our study, as depicted in figures 6-19. Each figure is organized with the problem statement at the top, the model's response on the lower left, and the annotators' comments to the lower right. For an extensive array of model responses and annotations, we invite readers to explore our data repository at: huggingface.co/datasets/mainlp/inferential_strategies.

C.1 Supposition Following

Figures from 6 to 8 demonstrate the application of *supposition following* by various models. For instance, Figure 6 presents LLaMA-2-70B's approach to problem 7, where the model supposes the absence of a blue marble in the box and logically infers the implications of this assumption to reach the valid conclusion. On the other hand, Figure 7 depicts Mistral-7B-Instruct's response to the same problem, where the model considers various combinations of marbles in the box, drawing immediate conclusions that follow from the premises at hand. However, it does not explore deeper ramifications of these suppositions, thereby failing to deduce the validity of the conclusion. This showcases a common behavior we observe in models that employ *supposition following* unsuccessfully. In Figure 8 the model approaches problem 9 by assuming the presence of an olive marble in the box, yet inferring disjointed intermediate conclusions that do not aid in solving the problem, thus failing to prove the logical validity of the problem.

C.2 Chain Construction

Figures 9 to 13 illustrate instances where models employ *chain construction* to navigate the problems of propositional logic. In Figure 9, LLaMA-2-70B adeptly forms a chain of conditional statements that bridge the antecedent of the conclusion to its consequent, effectively confirming the conclusion's logical validity. Conversely, Figure 10 depicts a logical chain in which LLaMA-2-70B erroneously concludes the nonexistence of a white marble based on the absence of a red marble, despite an exclusive disjunction linking the two. Despite this logical misstep, the model's final conclusion remains accurate, highlighting the discrepancy between the model's final answer and the soundness of its reasoning. In Figure 11, LLaMA-2-13B constructs a chain correctly linking the antecedent of the final conclusion to its consequent. Nonetheless, it overlooks the negation present in one of the conditionals, resulting in a compromised reasoning chain. Figure 12 presents a scenario where the model incorrectly attempts to validate an exclusive disjunction solely through a singular conditional sequence, a reasoning error not uncommon among human reasoners (Van der Henst et al., 2002). Lastly, Figure 13 highlights LLaMA-2-70B's engagement in the inverse fallacy, inferring $\neg$ W $\rightarrow$ $\neg$ G from the conditional W $\rightarrow$ G, mirroring a logical misjudgment frequently observed in human reasoning processes.

C.3 Compound Strategy

The *compound strategy* is illustrated in Figures 14 to 16. Figure 14 presents Mistral-7B-Instruct's approach to problem 9, where it infers a biconditional relationship between the purple and olive marble from the first two premises. On the other hand, Figure 15 shows LLaMA-2-70B's response to the same problem, formulating a sequence of compound inferences beyond the initial bicondi-

Model	Supposition Following	Chain Construction	Compound Conclusion	Concatenation Strategy	Symbolic Strategy	Correct Answer	Sound Reasoning
Zephyr-7B-β	60.0 ± 12.2	18.3 ± 6.2	10.0 ± 6.2	1.7 ± 3.3	20.0 ± 11.3	45.0 ± 15.5	25.0 ± 10.5
Mistral-7B-Instruct	35.0 ± 6.2	10.0 ± 3.3	35.0 ± 9.7	3.3 ± 4.1	8.3 ± 7.5	55.0 ± 10.0	25.0 ± 7.5
LLaMA-2-7B	20.0 ± 6.7	20.0 ± 15.5	6.7 ± 3.3	3.3 ± 4.1	1.7 ± 3.3	46.7 ± 6.7	0.0 ± 0.0
LLaMA-2-13B	28.3 ± 10.0	36.7 ± 12.5	6.7 ± 3.3	6.7 ± 6.2	0.0 ± 0.0	40.0 ± 8.2	15.0 ± 6.2
LLaMA-2-70B	45.0 ± 8.5	50.0 ± 7.5	3.3 ± 4.1	1.7 ± 3.3	6.7 ± 3.3	56.7 ± 6.2	31.7 ± 9.7

Table 4: Relative occurrences of inferential strategies employed by the different language models when solving the propositional problems. All values denote percentages averaged across 5 different random seeds with standard deviation. In addition, the percentages of correct final answers and sound reasoning are reported.

tional deduction, culminating in the correct final answer. Additionally, Figure 16 illustrates Mistral-7B-Instruct's approach to problem 8, in which the model initially generates compound conclusions derived from the problem statements, followed by *supposition following* to explore the implications that the absence of an olive marble might have. However, despite the model's sound reasoning, its final answer is incorrect.

C.4 Concatenation Strategy

Figure 17 demonstrates the *concatenation strategy*, where Mistral-7B-Instruct concatenates two intermediate deductions to form a single statement. It then uses the concatenated statement to infer the invalidity of the conclusion.

C.5 Symbolic Strategy

The *symbolic strategy* is exemplified in Figure 18, where LLaMA-2-70B employs a truth table to assess the conclusion's validity, albeit with errors leading to an incorrect result. Conversely, Figure 19 shows Mistral-7B-Instruct's application of *chain construction* followed by the *symbolic strategy*. The model makes false inferences while employing *chain construction*, and further errs in its validation through logical calculus.

Problem 7:

Statements:
1. There is a blue marble in the box if and only if there is a red marble in the box.
2. Either there is a red marble in the box or else there is a white marble in the box, but not both.
3. There is a white marble in the box if and only if there is a pink marble in the box.

Conclusion: If there is not a blue marble in the box then there is a pink marble in the box.

Response:
Sure, I'd be happy to help! Here's my reasoning process for this problem:

1. First, I'll start by analyzing the given statements:
Statement 1: "There is a blue marble in the box if and only if there is a red marble in the box"
This statement tells me that the presence of a blue marble in the box is dependent on the presence of a red marble in the box. If there's no red marble, there can't be a blue marble.

Statement 2: "Either there is a red marble in the box or else there is a white marble in the box, but not both"
This statement tells me that the presence of a red marble and a white marble in the box are mutually exclusive. If there's a red marble, there can't be a white marble, and vice versa.

Statement 3: "There is a white marble in the box if and only if there is a pink marble in the box"
This statement tells me that the presence of a white marble in the box is dependent on the presence of a pink marble in the box. If there's no pink marble, there can't be a white marble.

2. Now, let's consider the conclusion: "If there is not a blue marble in the box then there is a pink marble in the box"

I can start by assuming that there is no blue marble in the box.
According to statement 1, this means that there is no red marble in the box either.
Next, I can look at statement 2. Since there's no red marble, the alternative is that there's a white marble in the box. But wait, statement 3 says that if there's a white marble, there must be a pink marble too!
So, if there's no blue marble in the box, it means there's no red marble, which means there's a white marble, which means there's a pink marble

Therefore, the conclusion logically follows!
Conclusion: True

Annotator Comments:

Reformulation of statement 1.

Reformulation of statement 2.

Reformulation of statement 3.

Repetition of conclusion.

Supposition following (sound).
First consequence.
Second consequence.
Third consequence.

Conclusion (valid).

Final answer (correct).

Figure 6: The response (lower left box) of LLaMA-2-70B to problem 7 (top box) of the problem set, illustrating **supposition following**. After reformulating the statements of the problem at hand, the model starts to reason about the problem by assuming the absence of a blue marble in the box. Subsequently, it traces the consequences of that supposition, drawing intermediate conclusions about the presence or absence of certain marbles, until it formulates a final conclusion. In this example, the model correctly reasons about the validity of the conclusion.

Problem 7:

Statements:
1. There is a blue marble in the box if and only if there is a red marble in the box.
2. Either there is a red marble in the box or else there is a white marble in the box, but not both.
3. There is a white marble in the box if and only if there is a pink marble in the box.

Conclusion: If there is not a blue marble in the box then there is a pink marble in the box.

Response:
To determine if the conclusion logically follows from the given statements, let's first understand the given statements and the logical relationships between them:

1. Statement 1: "There is a blue marble in the box if and only if there is a red marble in the box."
- This statement is equivalent to saying that the presence of a blue marble and the presence of a red marble are mutually exclusive and exhaustive.

2. Statement 2: "Either there is a red marble in the box or else there is a white marble in the box, but not both."
- This statement tells us that there cannot be both a red and a white marble in the box.

3. Statement 3: "There is a white marble in the box if and only if there is a pink marble in the box."
- This statement is equivalent to saying that the presence of a white marble and the presence of a pink marble are mutually exclusive and exhaustive.

Now let's reason about the conclusion:
1. Conclusion: "If there is not a blue marble in the box then there is a pink marble in the box."

To determine if the conclusion logically follows from the given statements, we need to consider all possible cases:

Case 1: There is no blue marble in the box and there is a red marble.
- This case is not possible, as statement 1 tells us that a blue marble and a red marble cannot both be in the box.

Case 2: There is no blue marble in the box and there is a white marble.
- According to statement 3, if there is a white marble, then there must be a pink marble as well.

Case 3: There is no blue marble in the box and there is a pink marble.
- This case is not explicitly stated in the given statements, but it is possible based on the information provided.

Based on our analysis, we cannot definitively say that the conclusion logically follows from the given statements because there is a possible case (Case 1) that contradicts the conclusion. However, we have found that the conclusion is true in at least one possible case (Case 3).

Therefore, the conclusion "If there is not a blue marble in the box then there is a pink marble in the box" is not a logical consequence of the given statements, but it is true in at least one possible scenario.
Conclusion: False.

Annotator Comments:

Repetition of statement 1.

Reformulation of statement 1.

Repetition of statement 2.

Reformulation of statement 2.

Repetition of statement 3.

Reformulation of statement 3.

Repetition of conclusion.

Supposition following (invalid).

Supposition.

Supposition.

Supposition.

Conclusion (invalid).

Final answer (incorrect).

Figure 7: An exemplary model response of Mistral-7B-Instruct (lower left box) to problem 7 (top box) illustrating **supposition following**. The model successively assumes combinations of marbles in the box, and infers the immediate consequences from the premises provided. However, it does not extend its reasoning beyond the direct outcomes of each supposition, thereby failing to deduce the validity of the conclusion.

Problem 9:

Statements:
1. There is a purple marble in the box if and only if there is a blue marble in the box.
2. There is a blue marble in the box if and only if there is an olive marble in the box.
3. Either there is an olive marble in the box or else there is a red marble in the box, but not both.
4. Either there is a red marble in the box or else there is a green marble in the box, but not both.

Conclusion: If there is a purple marble in the box then there is a green marble in the box.

Response:	**Annotator Comments**:
Based on the given statements, here's my reasoning:	
1. First, let's analyze the given statements one by one:	
a) Statement 1: "There is a purple marble in the box if and only if there is a blue marble in the box"	Repetition of statement 1.
b) Statement 2: "There is a blue marble in the box if and only if there is an olive marble in the box"	Repetition of statement 2.
c) Statement 3: "Either there is an olive marble in the box or else there is a red marble in the box, but not both"	Repetition of statement 3.
d) Statement 4: "Either there is a red marble in the box or else there is a green marble in the box, but not both"	Repetition of statement 4.
2. Since we are trying to prove that "If there is a purple marble in the box then there is a green marble in the box", let's try to find a way to connect purple marbles with green marbles using the given statements.	Repetition of conclusion.
3. Unfortunately, there doesn't seem to be a direct connection between purple marbles and green marbles in the given statements.	
4. However, let's see if we can find any indirect connections. Based on Statement 1, we know that there is a purple marble in the box if and only if there is a blue marble in the box.	Reformulation of statement 1.
5. Now, let's look at Statement 2. We know that there is a blue marble in the box if and only if there is an olive marble in the box.	Reformulation of statement 2.
6. From Statement 3, we know that there cannot be both an olive marble and a red marble in the box. However, Statement 4 tells us that there cannot be both a red marble and a green marble in the box either.	Reformulation of statement 3. Reformulation of statement 4.
7. Let's consider the possibility that there is an olive marble in the box. If that's the case, then, according to Statement 2, there must be a blue marble in the box as well. 8. Now, using Statement 1, we know that if there is a blue marble in the box, then there must be a purple marble in the box as well. 9. However, Statement 3 tells us that if there is an olive marble in the box, then there cannot be a red marble in the box. But, Statement 4 tells us that if there is a red marble in the box, then there cannot be a green marble in the box.	Supposition following (sound).
10. Since we cannot have both an olive marble and a red marble in the box, and we cannot have both a red marble and a green marble in the box, it seems that the only possibility left is that there is an olive marble in the box, and therefore, a blue marble, a purple marble, but no red or green marbles.	Invalid inference.
11. However, the conclusion states that "If there is a purple marble in the box then there is a green marble in the box", which is not logically true based on the given statements.	Conclusion (invalid).
Conclusion: False.	Final answer (incorrect).

Figure 8: An exemplary model response of Mistral-7B-Instruct (lower left box) to problem 9 (top box) illustrating **supposition following**. The model supposes the presence of an olive marble in the box and traces the consequences of that supposition. However, it derives disjointed intermediate conclusions that do not aid in solving the problem, failing to solve the task at hand.

Problem 9:

Statements:
1. There is a purple marble in the box if and only if there is a blue marble in the box.
2. There is a blue marble in the box if and only if there is an olive marble in the box.
3. Either there is an olive marble in the box or else there is a red marble in the box, but not both.
4. Either there is a red marble in the box or else there is a green marble in the box, but not both.

Conclusion: If there is a purple marble in the box then there is a green marble in the box.

Response:
Sure, I'd be happy to help! Here's my reasoning process for this problem:

1. First, I'll start by analyzing the given statements:

Statement 1: "There is a purple marble in the box if and only if there is a blue marble in the box"
This statement tells me that the presence of a purple marble in the box is dependent on the presence of a blue marble in the box. If there is a blue marble, then there must be a purple marble, and vice versa.

Statement 2: "There is a blue marble in the box if and only if there is an olive marble in the box"
This statement tells me that the presence of a blue marble in the box is dependent on the presence of an olive marble in the box. If there is an olive marble, then there must be a blue marble, and vice versa.

Statement 3: "Either there is an olive marble in the box or else there is a red marble in the box, but not both"
This statement tells me that the presence of an olive marble and the presence of a red marble in the box are mutually exclusive. If there is an olive marble, then there cannot be a red marble, and vice versa.

Statement 4: "Either there is a red marble in the box or else there is a green marble in the box, but not both"
This statement tells me that the presence of a red marble and the presence of a green marble in the box are also mutually exclusive. If there is a red marble, then there cannot be a green marble, and vice versa.

2. Now, let's consider the conclusion: "If there is a purple marble in the box then there is a green marble in the box"

My reasoning process for this conclusion is as follows:

- If there is a purple marble in the box, then by Statement 1, there must be a blue marble in the box.
- If there is a blue marble in the box, then by Statement 2, there must be an olive marble in the box.
- If there is an olive marble in the box, then by Statement 3, there cannot be a red marble in the box.
- If there is no red marble in the box, then by Statement 4, there must be a green marble in the box.

Therefore, my conclusion is:
Conclusion: True

I hope that helps! Let me know if you have any questions or if you'd like me to explain my reasoning further.

Annotator Comments:

Reformulation of statement 1.

Reformulation of statement 2.

Reformulation of statement 3.

Reformulation of statement 4.

Repetition of conclusion.

Chain construction (sound).

$P \rightarrow B$

$B \rightarrow O$

$O \rightarrow \neg R$

$\neg R \rightarrow G$

Conclusion.
Final answer (correct).

Figure 9: The response (lower left box) of LLaMA-2-70B to problem 9 (top box) of the problem set, illustrating **chain construction**. The model correctly constructs a chain of conditionals leading from the antecedent of the final conclusion to its consequent.

Problem 7:

Statements:
1. There is a blue marble in the box if and only if there is a red marble in the box.
2. Either there is a red marble in the box or else there is a white marble in the box, but not both.
3. There is a white marble in the box if and only if there is a pink marble in the box.

Conclusion: If there is not a blue marble in the box then there is a pink marble in the box.

Response:	**Annotator Comments**:
Sure! Let's go through the statements and see if we can determine whether the conclusion logically follows.	
1. "There is a blue marble in the box if and only if there is a red marble in the box" This statement tells us that the presence of a blue marble in the box is dependent on the presence of a red marble. If there is a red marble, then there must be a blue marble, and if there is no red marble, then there cannot be a blue marble.	Reformulation of statement 1.
2. "Either there is a red marble in the box or else there is a white marble in the box, but not both" This statement tells us that the presence of a red marble and a white marble in the box are mutually exclusive. If there is a red marble, then there cannot be a white marble, and if there is a white marble, then there cannot be a red marble.	Reformulation of statement 2.
3. "There is a white marble in the box if and only if there is a pink marble in the box" This statement tells us that the presence of a white marble in the box is dependent on the presence of a pink marble. If there is a pink marble, then there must be a white marble, and if there is no pink marble, then there cannot be a white marble.	Reformulation of statement 3.
Now, let's consider the conclusion: "If there is not a blue marble in the box then there is a pink marble in the box"	Repetition of conclusion.
Let's break this down further:	Chain construction (invalid)
- If there is not a blue marble in the box, then we know that there is no red marble in the box (from statement 1).	$\neg$ B $\rightarrow$ $\neg$ R
- Since there is no red marble, we know that there cannot be a white marble in the box (from statement 2).	$\neg$ R $\rightarrow$ $\neg$ W (invalid)
- If there is no white marble in the box, then we know that there must be a pink marble in the box (from statement 3).	$\neg$ W $\rightarrow$ P (invalid)
Therefore, we can conclude that if there is not a blue marble in the box, then there must be a pink marble in the box.	Conclusion: $\neg$ B $\rightarrow$ P
Conclusion: True.	Final answer (correct).

Figure 10: The response (lower left box) of LLaMA-2-70B to problem 7 (top box) of the problem set, illustrating **chain construction**. The model constructs a chain of conditionals leading from the antecedent of the final conclusion to its consequent. However, it fails to understand the implication of the exclusive disjunction in the second statement of the problem description, leading to a faulty reasoning trace. Despite its invalid reasoning, the model's final answer is correct.

Problem 10:

Statements:
1. There is a brown marble in the box if and only if there is a yellow marble in the box.
2. Either there is a yellow marble in the box or else there is a green marble in the box, but not both.
3. There is a green marble in the box if and only if there is a purple marble in the box.
4. There is a purple marble in the box if and only if there is an olive marble in the box.

Conclusion: If there is a brown marble in the box then there is an olive marble in the box.

Response:	**Annotator Comments**:
Sure, I'd be happy to help with this experiment! Here's my reasoning process for each statement and the conclusion:	
1. There is a brown marble in the box if and only if there is a yellow marble in the box This statement means that if there is a brown marble in the box, then there must also be a yellow marble in the box. Conversely, if there is a yellow marble in the box, then there must also be a brown marble in the box.	Reformulation of statement 1.
2. Either there is a yellow marble in the box or else there is a green marble in the box, but not both This statement means that if there is a yellow marble in the box, then there cannot also be a green marble in the box. Similarly, if there is a green marble in the box, then there cannot also be a yellow marble in the box.	Reformulation of statement 2.
3. There is a green marble in the box if and only if there is a purple marble in the box This statement means that if there is a green marble in the box, then there must also be a purple marble in the box. Conversely, if there is a purple marble in the box, then there must also be a green marble in the box.	Reformulation of statement 3.
4. There is a purple marble in the box if and only if there is an olive marble in the box This statement means that if there is a purple marble in the box, then there must also be an olive marble in the box. Conversely, if there is an olive marble in the box, then there must also be a purple marble in the box.	Reformulation of statement 4.
Now, let's consider the conclusion: Conclusion: If there is a brown marble in the box then there is an olive marble in the box	Repetition of conclusion.
Based on the statements we've been given, we can logically deduce the conclusion as follows:	Chain construction (invalid).
1. If there is a brown marble in the box, then there must also be a yellow marble in the box (statement 1)	$B \rightarrow Y$
2. If there is a yellow marble in the box, then there cannot also be a green marble in the box (statement 2)	$Y \rightarrow \neg G$
3. If there is a green marble in the box, then there must also be a purple marble in the box (statement 3)	$G \rightarrow P$ (failure to consider negation)
4. If there is a purple marble in the box, then there must also be an olive marble in the box (statement 4)	$P \rightarrow O$
Therefore, if there is a brown marble in the box, then there must also be an olive marble in the box, which means the conclusion is true.	Conclusion (invalid).
Conclusion: True.	Final answer (incorrect).

Figure 11: The response (lower left box) of LLaMA-2-13B to problem 10 (top box) of the problem set, illustrating **chain construction**. The model constructs a chain of conditionals leading from the antecedent of the final conclusion to its consequent. However, it fails to account for the negation of the second conditional's consequent, leading to a faulty reasoning trace.

Problem 5:

Statements:
1. There is a purple marble in the box if and only if there is a yellow marble in the box.
2. There is a yellow marble in the box if and only if there is a blue marble in the box.
3. Either there is a blue marble in the box or else there is an orange marble in the box, but not both.

Conclusion: Either there is a purple marble in the box or else there is an orange marble in the box, but not both.

Response:	**Annotator Comments**:
Sure! Let's break this down step by step.	
1. "There is a purple marble in the box if and only if there is a yellow marble in the box" This statement implies that the presence of a purple marble in the box is dependent on the presence of a yellow marble in the box. If there is no yellow marble, there cannot be a purple marble.	Reformulation of statement 1.
2. "There is a yellow marble in the box if and only if there is a blue marble in the box" This statement implies that the presence of a yellow marble in the box is dependent on the presence of a blue marble in the box. If there is no blue marble, there cannot be a yellow marble.	Reformulation of statement 2.
3. "Either there is a blue marble in the box or else there is an orange marble in the box, but not both" This statement implies that the presence of a blue marble and an orange marble in the box are mutually exclusive. If there is a blue marble, there cannot be an orange marble, and vice versa.	Reformulation of statement 3.
Now, let's consider the conclusion: "Either there is a purple marble in the box or else there is an orange marble in the box, but not both"	Repetition of conclusion. Chain construction (sound).
Based on the first two statements, we know that if there is a purple marble, there must be a yellow marble, and if there is a yellow marble, there must be a blue marble.	$P \rightarrow Y$ $Y \rightarrow B$
Therefore, if there is a purple marble, there must be a blue marble.	$P \rightarrow B$
Now, let's consider the third statement. If there is a blue marble, there cannot be an orange marble.	$B \rightarrow \neg O$
Therefore, if there is a purple marble (which implies a blue marble), there cannot be an orange marble.	$P \rightarrow \neg O$
Conclusion: True. The conclusion logically follows from the given statements.	Final answer (correct). Conclusion (invalid).

Figure 12: The response (lower left box) of LLaMA-2-70B to problem 5 (top box) of the problem set, illustrating **chain construction**. The model constructs a chain of conditionals proving one case of the exclusive disjunction. However, it fails to account for the other conditional case, i.e. $\neg P \rightarrow O$, therefore failing to prove the logical validity of the conclusion.

Problem 12:

Statements:
1. There is a blue marble in the box if and only if there is a brown marble in the box.
2. There is a white marble in the box if and only if there is a green marble in the box.
3. Either there is a brown marble in the box or else there is a white marble in the box, but not both.
4. There is a green marble in the box if and only if there is a purple marble in the box

Conclusion: If there is a blue marble in the box then there is a purple marble in the box.

Response:	**Annotator Comments**:
Sure! Let's break down the problem and think through each step.	
1. First, we have the statement "There is a blue marble in the box if and only if there is a brown marble in the box." This means that if there is a blue marble in the box, then there must be a brown marble in the box, and vice versa.	Reformulation of statement 1.
2. Next, we have the statement "There is a white marble in the box if and only if there is a green marble in the box." This means that if there is a white marble in the box, then there must be a green marble in the box, and vice versa.	Reformulation of statement 2.
3. The third statement is "Either there is a brown marble in the box or else there is a white marble in the box, but not both." This means that if there is a brown marble in the box, then there cannot be a white marble in the box, and vice versa.	Reformulation of statement 3.
4. The fourth statement is "There is a green marble in the box if and only if there is a purple marble in the box." This means that if there is a green marble in the box, then there must be a purple marble in the box, and vice versa.	Reformulation of statement 4.
5. Now, let's consider the conclusion: "If there is a blue marble in the box then there is a purple marble in the box."	Repetition of conclusion.
Let's think about how we can use the given statements to determine whether this conclusion is true or false.	Chain construction (invalid).
1. We know that if there is a blue marble in the box, then there must be a brown marble in the box (from statement 1).	Blue $\rightarrow$ Brown
2. We also know that if there is a brown marble in the box, then there cannot be a white marble in the box (from statement 3).	Brown $\rightarrow \neg$ W
3. Therefore, if there is a blue marble in the box, then there cannot be a white marble in the box.	Blue $\rightarrow \neg$ W
4. We also know that if there is a white marble in the box, then there must be a green marble in the box (from statement 2).	W $\rightarrow$ G
5. Therefore, if there is a blue marble in the box, then there cannot be a green marble in the box.	Blue $\rightarrow \neg$ G (fallacy of the inverse).
6. However, we know that if there is a green marble in the box, then there must be a purple marble in the box (from statement 4).	G $\rightarrow$ P
7. Therefore, if there is a blue marble in the box, then there must be a purple marble in the box.	Blue $\rightarrow$ P (invalid)
Conclusion: True.	Final answer (correct).
In summary, we have shown that if there is a blue marble in the box, then there must be a purple marble in the box, using the given statements. Therefore, the conclusion logically follows from the statements.	Conclusion (invalid).

Figure 13: The response (lower left box) of LLaMA-2-70B to problem 12 (top box) of the problem set, illustrating **chain construction**. The model constructs a chain of conditionals leading from the antecedent of the final conclusion to its consequent. However, it makes a series of mistakes when constructing the chain of conditionals. For instance, it infers the absence of the green marble by denying the presence of the white marble, i.e. Blue $\rightarrow \neg$ W; W $\rightarrow$ G; therefore Blue $\rightarrow \neg$ G by assuming that $\neg$ W $\rightarrow \neg$ G, which is a common logical error known as the *fallacy of the inverse*.

Problem 9:

Statements:
1. There is a purple marble in the box if and only if there is a blue marble in the box.
2. There is a blue marble in the box if and only if there is an olive marble in the box.
3. Either there is an olive marble in the box or else there is a red marble in the box, but not both.
4. Either there is a red marble in the box or else there is a green marble in the box, but not both.

Conclusion: If there is a purple marble in the box then there is a green marble in the box.

Response:	**Annotator Comments**:
First, let's analyze the given statements:	
1. Statement 1: There is a purple marble in the box if and only if there is a blue marble in the box	Reformulation of statement 1.
2. Statement 2: There is a blue marble in the box if and only if there is an olive marble in the box	Reformulation of statement 2.
3. Statement 3: Either there is an olive marble in the box or else there is a red marble in the box, but not both	Reformulation of statement 3.
4. Statement 4: Either there is a red marble in the box or else there is a green marble in the box, but not both	Reformulation of statement 4.
Let's examine the relationship between the statements:	Compound Strategy (sound).
From Statement 1 and Statement 2, we know that there is a purple marble in the box if and only if there is a blue marble in the box, and there is a blue marble in the box if and only if there is an olive marble in the box. Therefore, we can conclude that there is a purple marble in the box if and only if there is an olive marble in the box.	List biconditional statements: $P \leftrightarrow B$ $B \leftrightarrow O$ Draw compound conclusion: $P \leftrightarrow O$
However, the conclusion states that if there is a purple marble in the box, then there is a green marble in the box. This is not logically following from the given statements.	Conclusion (invalid).
Conclusion: False.	Final answer (incorrect).

Figure 14: The response (lower left box) of Mistral-7B-Instruct to problem 9 (top box) of the problem set, illustrating the **compound strategy**. Based on the first two premises of the problem description, the model draws a compound conclusion, establishing equivalence between the purple and olive marble in the box. However, Mistral-7B-Instruct fails to draw additional intermediate conclusions that would be required to deduce the logical validity of the conclusion in the problem statement.

Problem 9:

Statements:
1. There is a purple marble in the box if and only if there is a blue marble in the box.
2. There is a blue marble in the box if and only if there is an olive marble in the box.
3. Either there is an olive marble in the box or else there is a red marble in the box, but not both.
4. Either there is a red marble in the box or else there is a green marble in the box, but not both.

Conclusion: If there is a purple marble in the box then there is a green marble in the box.

Response:	**Annotator Comments**:
Sure! Let's break it down step by step.	
1. "There is a purple marble in the box if and only if there is a blue marble in the box" This statement implies that the presence of a purple marble in the box is dependent on the presence of a blue marble in the box. If there is no blue marble, there cannot be a purple marble.	Reformulation of statement 1.
2. "There is a blue marble in the box if and only if there is an olive marble in the box" This statement implies that the presence of a blue marble in the box is dependent on the presence of an olive marble in the box. If there is no olive marble, there cannot be a blue marble.	Reformulation of statement 2.
3. "Either there is an olive marble in the box or else there is a red marble in the box, but not both" This statement implies that the presence of an olive marble and a red marble in the box are mutually exclusive. If there is an olive marble, there cannot be a red marble, and vice versa.	Reformulation of statement 3.
4. "Either there is a red marble in the box or else there is a green marble in the box, but not both" This statement implies that the presence of a red marble and a green marble in the box are also mutually exclusive. If there is a red marble, there cannot be a green marble, and vice versa.	Reformulation of statement 4.
Now, let's evaluate the conclusion: Conclusion: If there is a purple marble in the box then there is a green marble in the box	Repetition of conclusion.
Based on the given statements, we know that the presence of a purple marble is dependent on the presence of a blue marble, and the presence of a blue marble is dependent on the presence of an olive marble. Therefore, if there is a purple marble in the box, it means that there is also a blue marble and an olive marble in the box.	Compound Strategy (sound). $P \leftrightarrow B$ $B \leftrightarrow O$ Draw compound conclusion: $P \leftrightarrow O$
Using the third statement, we know that the presence of an olive marble means that there cannot be a red marble in the box. Therefore, if there is a purple marble in the box, there cannot be a red marble.	Derive conditional (statement 3): $O \rightarrow \neg R$ Draw compound conclusion.
Using the fourth statement, we know that the absence of a red marble means that there must be a green marble in the box. Therefore, if there is a purple marble in the box, there must be a green marble in the box as well.	Derive conditional (statement 4): $\neg R \rightarrow G$ Draw compound conclusion.
Conclusion: True So, the conclusion logically follows from the given statements.	Final answer (correct). Conclusion (valid).

Figure 15: The response (lower left box) of LLaMA-2-70B to problem 9 (top box) of the problem set, illustrating the **compound strategy**. The model draws a series of compound conclusions to deduce the logical validity of the conclusion in the problem statement.

Problem 8:

Statements:
1. Either there is an olive marble in the box or else there is a brown marble in the box, but not both.
2. There is a brown marble in the box if and only if there is a gray marble in the box.
3. Either there is a gray marble in the box or else there is a maroon marble in the box, but not both.

Conclusion: If there is not an olive marble in the box then there is a maroon marble in the box.

Response:
To determine if the conclusion logically follows from the given statements, we need to use the principles of logical reasoning. Let's analyze the statements one by one and then consider the conclusion.

1. Statement 1: Either there is an olive marble in the box or else there is a brown marble in the box, but not both

2. Statement 2: There is a brown marble in the box if and only if there is a gray marble in the box

3. Statement 3: Either there is a gray marble in the box or else there is a maroon marble in the box, but not both

First, let's understand the implications of Statement 2. It states that if there is a brown marble in the box, then there is a gray marble in the box, and conversely, if there is a gray marble in the box, then there is a brown marble in the box. This is an if and only if statement, meaning both conditions are equivalent.

Now, let's consider Statement 1. It states that there is either an olive marble or a brown marble in the box, but not both.
Since Statement 2 tells us that a brown marble implies a gray marble,
it follows that if there is no olive marble (i.e., there is a brown marble), then there must be a gray marble.
However, Statement 3 tells us that there is either a gray marble or a maroon marble in the box, but not both.
Therefore, if there is a gray marble, then there cannot be a maroon marble.

So, based on the given statements, if there is no olive marble in the box, then there is a brown marble (from Statement 1 and Statement 2), which implies a gray marble (from Statement 2), and since there is no olive marble and there is a gray marble, there cannot be a maroon marble (from Statement 3).

Therefore, the conclusion follows logically from the given statements.
Conclusion: True.

Annotator Comments:

Repetition of statement 1.

Repetition of statement 2.

Repetition of statement 3.

Reformulation of statement 2.

Compound strategy (valid).
$O \oplus B$
$B \leftrightarrow G$
Draw compound conclusion:
$\neg O \rightarrow G$
$G \oplus M$
Draw compound conclusion:
$\neg O \rightarrow \neg M$

Supposition Following (valid).
Tracing consequences.

Conclusion (invalid).
Final answer (incorrect).

Figure 16: The response (lower left box) of Mistral-7B-Instruct to problem 8 (top box) of the problem set, illustrating the **compound strategy** and **supposition following**. Based on the first two premises of the problem description, the model first draws a compound conclusion, establishing that a gray marble follows from the absence of an olive marble. Subsequently, it uses this intermediate conclusion, together with the third premise, to draw another compound conclusion about the absence of the maroon marble. The model then switches to *supposition following*, tracing the consequences of the absence of the olive marble, inferring the final conclusion that there cannot be a maroon marble. However, despite the model's correct reasoning, it deduces the wrong answer: "True".

Problem 6:

Statements:
1. There is a gray marble in the box if and only if there is a yellow marble in the box.
2. Either there is a yellow marble in the box or else there is an olive marble in the box, but not both.
3. There is an olive marble in the box if and only if there is a black marble in the box.

Conclusion: If there is a yellow marble in the box then there is a black marble in the box.

Response:	**Annotator Comments**:
Based on the given statements, here is my reasoning:	
1. The first statement tells us that there is a gray marble in the box if and only if there is a yellow marble in the box.	Reformulation of statement 1.
2. The second statement tells us that either there is a yellow marble in the box or there is an olive marble in the box, but not both.	Reformulation of statement 2.
3. The third statement tells us that there is an olive marble in the box if and only if there is a black marble in the box.	Reformulation of statement 3.
4. From the second statement, we know that if there is a yellow marble in the box, then there cannot be an olive marble in the box.	Conditional from statement 2. $Y \rightarrow \neg O$
5. From the third statement, we know that if there is an olive marble in the box, then there is a black marble in the box.	Conditional from statement 3. $O \rightarrow B$
6. Since the second statement tells us that there cannot be both a yellow marble and an olive marble in the box, it follows that if there is a yellow marble in the box, then there cannot be an olive marble or a black marble in the box (because of the if and only if relationship in the third statement).	Concatenation of statements (sound). $Y \rightarrow \neg (O \leftrightarrow B)$
7. Therefore, the conclusion does not logically follow from the given statements.	Conclusion (valid).
Conclusion: False.	Final answer (correct).

Figure 17: The response (lower left box) of Mistral-7B-Instruct to problem 6 (top box) of the problem set, illustrating the **concatenation strategy**. Mistral-7B-Instruct concatenates the intermediate conditional conclusion ($Y \rightarrow \neg O$) and the third premise of the problem statement ($O \leftrightarrow B$) to form the concatenated conclusion $Y \rightarrow \neg (O \leftrightarrow B)$. Based on that conclusion, the model infers that the conclusion in the problem statement does not logically follow from the premises at hand.

Problem 3:

Statements:
1. There is a green marble in the box if and only if there is a purple marble in the box
2. If there is a purple marble in the box then there is a gray marble in the box.
3. Either there is a gray marble in the box or else there is a yellow marble in the box, but not both.

Conclusion: Either there is a green marble in the box or else there is a yellow marble in the box, but not both.

Response:
Sure, I'd be happy to help! Here's my reasoning process for this problem:

1. First, I'll start by analyzing the given statements:
- Statement 1: "There is a green marble in the box if and only if there is a purple marble in the box"
- This statement establishes a direct relationship between the presence of a green marble and a purple marble in the box.

- Statement 2: "If there is a purple marble in the box then there is a gray marble in the box"
- This statement establishes a conditional relationship between the presence of a purple marble and a gray marble in the box.

- Statement 3: "Either there is a gray marble in the box or else there is a yellow marble in the box, but not both"
- This statement establishes an exclusive OR relationship between the presence of a gray marble and a yellow marble in the box.

2. Now, let's consider the conclusion: "Either there is a green marble in the box or else there is a yellow marble in the box, but not both"

3. To determine whether the conclusion logically follows from the statements, I'll use a truth table analysis:

	Statement 1	Statement 2	Statement 3	Conclusion
1	T	T	T	T
2	T	T	F	F
3	T	F	T	F
4	F	T	T	F
5	F	T	F	F
6	F	F	T	T
7	F	F	F	F

As we can see from the truth table, the conclusion "Either there is a green marble in the box or else there is a yellow marble in the box, but not both" is true in all cases where Statement 1 and Statement 2 are true, and false in all cases where Statement 3 is false.

Since Statement 1 and Statement 2 are always true, and Statement 3 is true in some cases, the conclusion logically follows from the statements.
Conclusion: True

Annotator Comments:

Reformulation of statement 1.

Reformulation of statement 2.

Reformulation of statement 3.

Repetition of conclusion.

Symbolic Strategy (invalid).

Construct truth table.

Summarize truth table.

Conclusion (invalid).

Final answer (correct).

Figure 18: The response (lower left box) of LLaMA-2-70B to problem 3 (top) of the problem set, illustrating the **symbolic strategy**. The model constructs a truth table to infer the validity of the conclusion given in the problem statement. However, the model produces errors in the truth table, resulting in flawed reasoning.

Problem 6:

Statements:
1. There is a gray marble in the box if and only if there is a yellow marble in the box.
2. Either there is a yellow marble in the box or else there is an olive marble in the box, but not both.
3. There is an olive marble in the box if and only if there is a black marble in the box.

Conclusion: If there is a yellow marble in the box then there is a black marble in the box.

Response:	**Annotator Comments**:
First, we need to understand the relationships between the marbles and the statements. Here's a summary:	
- There is a gray marble in the box if and only if there is a yellow marble in the box - This statement creates a connection between the gray marble and the yellow marble. If one is present, the other is as well.	Reformulation of statement 1.
- Either there is a yellow marble in the box or else there is an olive marble in the box, but not both - This statement introduces two possible scenarios. Either the yellow marble is present, or the olive marble is present, but not both.	Reformulation of statement 2.
- There is an olive marble in the box if and only if there is a black marble in the box - This statement creates another connection, but this time between the olive marble and the black marble. If one is present, the other is as well.	Reformulation of statement 3.
Now, let's see if the conclusion follows logically from these statements. - If there is a yellow marble in the box then there is a black marble in the box	Repetition of conclusion.
- To prove this conclusion, we'll use the second statement to eliminate the possibility of the olive marble being present. Then, we'll use the third statement to connect the presence of the yellow marble with the presence of the black marble.	Chain construction (invalid).
- If there is a yellow marble in the box, then by the second statement, the olive marble cannot also be present.	$Y \to \neg O$
- If the olive marble is not present, then by the third statement, the black marble must be present.	$\neg O \to B$ (invalid inference).
- Therefore, if there is a yellow marble in the box, then there is a black marble in the box.	$Y \to B$
Conclusion: True.	Final answer (incorrect).
Explanation: - Statement 1: Gray marble ↔ Yellow marble - Statement 2: Yellow marble ∨ !Olive marble - Statement 3: Olive marble ↔ Black marble - Conclusion: Yellow marble → Black marble	Symbolic Strategy (invalid). Translate statement 1. Translate statement 2. (invalid) Translate statement 3. Translate conclusion.
Reasoning steps: 1. Given: Yellow marble → Black marble (conclusion) 2. Given: Olive marble ↔ Black marble (statement 3) 3. Given: Yellow marble ∨ !Olive marble (statement 2) 4. Eliminate olive marble: !Olive marble (from statement 2) 5. Given: Gray marble ↔ Yellow marble (statement 1) 6. Eliminate gray marble: !Gray marble (from statement 1) 7. Combine 4 and 6: Yellow marble → !Gray marble (from statements 1 and 2) 8. Combine 7 and 3: Yellow marble → Black marble (from statements 2, 3, and 7) 9. Conclusion: True.	 Repeat formal conclusion. Repeat formal statement 3. Repeat formal statement 2. Inference. Repeat formal statement 1. Invalid inference. Final answer (incorrect).
These steps show that the conclusion follows logically from the given statements.	Conclusion (invalid).

Figure 19: The response (lower left box) of Zephyr-7B-β to problem 6 (top box) of the problem set, illustrating **chain construction** and the **symbolic strategy**. The model first constructs a chain of conditionals to prove the validity of the conclusion, linking relevant entities in premise two and three of the problem statement. Subsequently, the model *"explains"* its reasoning by employing the symbolic strategy, converting statements into formal logic and operating on them. Note that the model makes several logical errors on its way to prove the logical validity of the final conclusion.