Conditional and Modal Reasoning in Large Language Models

Wesley H. Holliday[1] **Matthew Mandelkern**[2] **Cedegao E. Zhang**[3]
[1]University of California, Berkeley
[2]New York University
[3]Massachusetts Institute of Technology
wesholliday@berkeley.edu, mandelkern@nyu.edu, cedzhang@mit.edu

Abstract

The reasoning abilities of large language models (LLMs) are the topic of a growing body of research in AI and cognitive science. In this paper, we probe the extent to which twenty-nine LLMs are able to distinguish logically correct inferences from logically fallacious ones. We focus on inference patterns involving conditionals (e.g., '*If* Ann has a queen, *then* Bob has a jack') and epistemic modals (e.g., 'Ann *might* have an ace', 'Bob *must* have a king'). These inferences have been of special interest to logicians, philosophers, and linguists, since they play a central role in the fundamental human ability to reason about distal possibilities. Assessing LLMs on these inferences is thus highly relevant to the question of how much the reasoning abilities of LLMs match those of humans. All the LLMs we tested make some basic mistakes with conditionals or modals, though zero-shot chain-of-thought prompting helps them make fewer mistakes. Even the best performing LLMs make basic errors in modal reasoning, display logically inconsistent judgments across inference patterns involving epistemic modals and conditionals, and give answers about complex conditional inferences that do not match reported human judgments. These results highlight gaps in basic logical reasoning in today's LLMs.

1 Introduction

One of the most distinctive human cognitive abilities is the ability to think about what follows *if* something is the case—conditional thinking—and about what *might* or *must* be the case—modal thinking (Evans and Over, 2004; Portner, 2009). Such reasoning about distal possibilities is crucial to the human capacity for *planning* (we try to choose the action that *would* bring about the best effects *if* we were to take it (Gibbard and Harper, 1981)), *causal reasoning* (C causes E if E *wouldn't* have happened *if* C hadn't (Lewis, 1973a; Beller and Gerstenberg, 2023)), *retroactive evaluation*, and more.

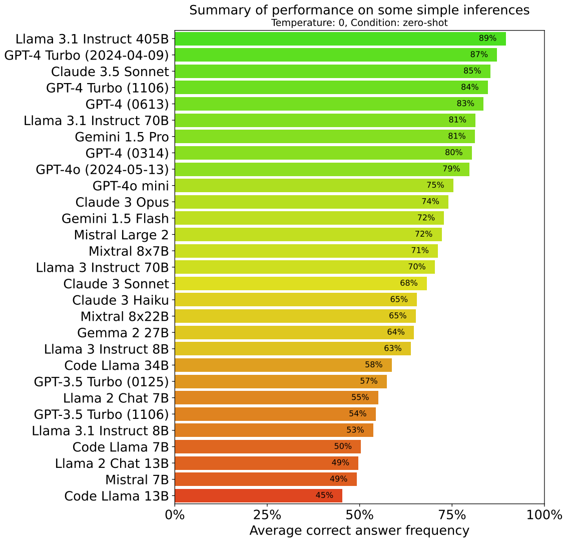


Figure 1: Summary of performance on the uncontroversial logical inference patterns discussed in § 4. Guessing accuracy is 50%. Larger models generally perform better, and most models show clear weakness at this task.

Conditional and modal language has thus been a central focus of philosophers (e.g., Stalnaker, 1968; Lewis, 1973b; Khoo, 2022), linguists (e.g., Kratzer, 2012; Portner, 2009), and logicians (e.g., Kripke, 1963; Stalnaker and Thomason, 1970; van Benthem, 2023), as well as an interest of computer scientists (e.g., Friedman and Halpern, 1994; Fagin et al., 1995), leading to a variety of sophisticated models of conditional and modal reasoning (Egré and Rott, 2021; Garson, 2024).

With the rapid recent development of large language models (LLMs) that at least superficially resemble human speakers and reasoners in many respects (Huang and Chang, 2022; Wei et al., 2022; Bubeck et al., 2023; Zhao et al., 2023), a natural question to ask is to what extent LLMs have mastered conditional and modal reasoning. In this paper, we begin to tackle this problem from the perspective of philosophers and logicians, probing the degree to which different LLMs have mastered the

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3800–3821
November 12-16, 2024 ©2024 Association for Computational Linguistics

logical inference patterns characteristic of conditional and modal reasoning. For example, consider the pattern known as *Modus Tollens* (MT): 'If p, then q. Not q. Therefore, not p'. We tested whether LLMs draw inferences in accord with this pattern by prompting them with many instances of the pattern, as in:

> **User prompt**: From 'If Alex finished the race, then Chris finished the race' together with 'Chris did not finish the race', can we infer 'Alex did not finish the race'? (**System prompt**: Answer only with 'yes' or 'no' and nothing else.)
>
> **GPT-4**: yes. **Mistral 7B**: no. Etc.

We then gave other instances of the pattern to each LLM, asssessing their performance in terms of accuracy on the pattern of inference. Figure 1 summarizes performance across several inference patterns to be discussed. We also compared performance on the zero-shot condition shown above with few-shot and chain-of-thought conditions (Table 2).

After providing some background in § 2 and detailing our experimental setup in § 3, we discuss results for a number of inference patterns in § 4. We find that among the LLMs tested, all models commit basic fallacies in reasoning with modals and conditionals. Even the best performing models display logically inconsistent judgments across certain inference patterns involving modals and conditionals. And almost all models give answers to certain complex conditional inferences that do not match reported human judgments. We also show that models' performance on our reasoning tasks is highly correlated with that of Chatbot Arena Elo ratings (Chiang et al., 2024), MMLU (Hendrycks et al., 2020), and GSM8K (Cobbe et al., 2021), supporting the hypothesis that logical reasoning abilities are predictive of general model capabilities and performance on downstream tasks. In sum, our main contributions in this paper are:

- Emphasizing the importance and nuances of reasoning about conditionals and modals, grounded in up-to-date evidence and theories from the relevant literature.
- Proposing a focused, novel benchmark that tests LLMs' ability to engage in logical reasoning with conditionals and modals.
- Reporting the performance of a large set of LLMs in different prompting settings and identifying some of their gaps and undesirable behaviors in basic logical reasoning.

2 Background and related work

Our goal is to apply methodologies from the philosophical, logical, and linguistic literature on conditionals and modals to the study of LLMs.

2.1 Logical inference

First and most generally, we draw on a philosophical understanding of what a *logical inference* is. Logical inferences are those inferences that are valid just in virtue of the meaning of *logical* words like 'and', 'or', 'not', 'if', 'must', 'might', and so on. That is, a logically valid inference is one whose conclusion is always true when its premises are, *no matter how the non-logical words in the premises and conclusion are understood* (Tarski, 1936).

This contrasts with more colloquial uses of 'logical reasoning' that are current in the literature on LLMs, where 'logical reasoning' is often used for reasoning in general, involving inferential leaps of various kinds that go beyond deductive inference proper (this holds for many of the tasks studied in Xu et al. (2023); Chen et al. (2023); Huang and Chang (2023); Liu et al. (2021) and nearly all the BigBench tasks (BIG-bench authors, 2023) categorized under the keyword 'logical reasoning'). For instance, in the logicians' sense, the inference 'A is to the left of B, hence B is to the right of A' is not *logically* valid, since its correctness depends on the meaning of the non-logical words 'left' and 'right'. By contrast, 'A is to the left of B, hence something is to the left of B' is logically valid, since its correctness relies only on the meaning of the logical word 'something'. While studying content-based reasoning in LLMs is obviously of great interest, we believe it is also of fundamental interest to study purely logical reasoning in LLMs, since such reasoning is plausibly part of the backbone of human inference and knowledge of meanings.

Regarding purely logical reasoning, a series of benchmarks have been created in recent years (Tafjord et al., 2020; Tian et al., 2021; Han et al., 2022; Saparov and He, 2023; Saparov et al., 2023), and various strategies have been proposed to solve some of them (Creswell et al., 2023; Kazemi et al., 2023; Olausson et al., 2023; Pan et al., 2023; Poesia et al., 2023; Ye et al., 2023). Those studies (with the exceptions of Tafjord et al. 2020; Creswell et al. 2023) primarily focus on *multi-step* reasoning, where a proof is required from premises to the hypothesis. Here we target single-step inference patterns, which we treat as more fundamental. Indeed, the inability to recognize the basic inference patterns we study here could provide further explanations of failures on multi-step reasoning problems. Most importantly, none of the work above studies modal operators, conditional operators, and their interactions, our novel focus in this paper.

2.2 Modals and conditionals

Here we draw specifically on the logical and philosophical literature on *modals* and *conditionals*. Enormous progress has been made in the last half century on both topics. First, modal operators (like 'must' and 'might') have been successfully modeled as *quantifiers over possible worlds* (Kripke, 1963; Kratzer, 1981). That is, just as 'Every boy is sitting' quantifies universally over all boys (in a given domain), 'It must be raining' quantifies over all possible worlds (in a given domain—in this case, an *epistemic* domain) and says that it is raining in all of them; and just as 'Some boy is sitting' quantifies existentially over boys, 'It might be raining' says that it is raining in some epistemically possible world. Similarly, a deontic modal like 'may' can be interpreted as quantifying over a domain of deontically possible worlds, so that 'You may eat a cookie' is true just in case you eat a cookie in some deontically possible world, i.e., one where all the actual deontic requirements are satisfied. Likewise, 'must', on its deontic interpretation (as in 'You must eat this cookie') quantifies *universally* over deontically possible worlds, and says that you eat cookies in all of them.

This interpretation of modals yields corresponding *logics* of modal language, with the details depending on how the domain of possible worlds is obtained (and the interpretation of the other connectives and operators with which modals interact).

Conditional operators have likewise been analyzed with possible worlds semantics. In classical logic, 'if p, then q' is treated as the *material conditional*, which is true whenever p is false or q is true. However, it is almost universally accepted by philosophers, linguists, and logicians that this treatment is a very poor approximation to the actual meaning of 'if' in natural language. For instance, on the material analysis of 'if', 'No student will fail if she studies hard' would entail 'Every student will study hard', which obviously does not follow. Likewise, if the material analysis were correct, then the probability of 'if p, then q' would go up as the probability of p goes down, but this is wrong. Consider a fair coin. The probability that the coin will land heads if it is flipped is intuitively .5, and it is intuitively probabilistically *independent* of whether the coin is flipped. That is, finding out that the coin probably will not be flipped does not make it any more likely that if it is flipped, it will land heads (Douven and Verbrugge, 2013). Edgington (1995) provides a battery of widely accepted further arguments against the material analysis.

These points are worth emphasizing, since although the material analysis is almost universally rejected by theorists of the conditional, it is still assumed in much existing work testing the logical capacities of humans and LLMs, in both cognitive science and artificial intelligence (e.g., in the recent Wan et al. 2024, which treats the material analysis as one of the benchmarks of *correct* reasoning with conditionals). This is a serious blindspot, since failing to reason in accord with the material conditional may be logically *correct*; and, conversely, reasoning in accord with the material conditional may be a serious logical mistake.

The most popular alternative treats 'if p, then q' as a restricted modal operator, which says that q is true in all p-worlds (in a given domain). Just as for modals, this yields corresponding logics, with the details again depending on assumptions about which p-worlds are in the domain, together with the interpretation of other connectives (Stalnaker, 1968; Lewis, 1973b; Egré and Rott, 2021).

Although the material analysis is almost universally rejected, there is ongoing controversy about the correct logic of conditionals and modals. We have chosen a wide range of inference patterns to test: in many of these cases there is (near) universal agreement about whether the inference pattern is valid. In other cases, there is less agreement about whether the pattern is truly valid, but even in those cases, there is for the most part agreement about whether typical human reasoners are inclined to draw the inference, and the remaining controversy is about how to model those patterns (as genuine (in)validities or the result of systematic shifts in interpretation). We do not aim to take a position in these complex debates here but rather to compare the behavior of LLMs to widely reported human inferential dispositions. In future work, we plan to compare the behavior of LLMs with human subjects (compare the methodology of Pavlick and Kwiatkowski, 2019; Dasgupta et al., 2022; Webson et al., 2023); here we compare LLMs against expert claims about inference from the literature in philosophy, logic, and semantics.

2.3 Natural language inference

Our task format and evaluation method is similar to the one used in the natural language inference (NLI) paradigm (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2019), which has a rich and long tradition (Katz, 1972; Condoravdi et al., 2003; van Benthem, 2008; MacCartney and Manning, 2009; Dagan et al., 2010). There, a problem comes with a premise P and a hypothesis H, and the goal is to decide whether the premise entails, contradicts, or is neutral with respect to H. The notion of entailment is typically based on common sense, whereas in this work we exclusively study logical entailment in the sense specified above.

In sum, our approach differs from previous work on LLMs in two central ways: (i) we focus on one-step *logical* inference, in the austere philosophical sense, rather than common-sense reasoning in general, differing from most extant benchmarks; (ii) we bring sophisticated approaches to the logic of conditionals and modals from philosophy, linguistics, and logic, yielding new ways to assess how closely LLMs match human reasoning in this key domain. In particular, in contrast to the work on logical reasoning cited above, we go beyond propositional/predicate logic to incorporate more realistic approaches to the logic of conditionals and modals, which to our knowledge has not been explored.

3 Experiments

3.1 Models

We tested the logical inference judgments of the 29 LLMs listed in Figure 1, including both open and closed ones, cited in Appendix A. The Anthropic, Google, Mistral, and OpenAI models were run through their respective APIs. Open-weight models were run through cloud providers (together.ai and fireworks.ai). All experiments cost ∼\$3,000 for API calls. All code and data for the experiments are available at github.com/wesholliday/llm-logic, which also includes data for OpenAI's o1 models (released after this paper was submitted).

3.2 Data

For our experiments, we created a bank of questions for dozens of inference patterns, allowing us to individually probe LLMs' inferential capacities with respect to each inference pattern. For each of these, we began by handcrafting a paradigm instance. E.g., for Modus Tollens (MT) (see Table 1), our paradigm instance was: "From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Sue was not at the wedding', can we infer 'Mary was not at the wedding'?" With the help other other LLMs (mostly Claude 2, supplemented by GitHub Copilot and Devin), we created 19 additional instances of the pattern, e.g., "From 'If the alien visited Mars, then the robot visited Jupiter' together with 'The robot did not visit Jupiter', can we infer 'The alien did not visit Mars'?" To guard against an LLM judging an inference based on world knowledge rather than the inference's logical form, we also created 20 instances with *nonsense predicates*, e.g., "From 'If the flugel was blimmed, then the flugel was zargled' together with 'The flugel was not zargled', can we infer 'The flugel was not blimmed'?" We denote the version of an inference with nonsense predicates with an 'x' at the end of its name, e.g., MTx. We also tested order effects for all inferences with two premises by switching the order of premises in an 'o' version. Using nonsense predicates and switched premise order yields an 'ox' version that we also tested. We reviewed all LLM-generated stimuli and made necessary adjustments by hand.

3.3 Evaluation

For each of the 20 instances of an inference pattern question and each LLM, we posed the instance to the LLM with temperature 0 and then 1, along with a prompt for either zero-shot, few-shot, or zero-shot chain-of-thought (Kojima et al., 2022) conditions (see Appendix B). For temperature $= 1$, we asked the question repeatedly to get an empirical frequency for 'yes' and for 'no'.[1]

To assess the sufficiency of using 20 instances of each inference pattern, we looked at the Pearson correlation coefficients between the yes-frequency of model responses to the 20 instances of an inference with nonsensical predicates (like MTx) and the yes-frequency of model responses to the 20 instances with sensical predicates (like MT). The

[1] We used the following stopping rule: if for a particular question the LLM answered 'yes' 10 times in a row or 'no' 10 times in a row, then we proceeded to the next question; otherwise we posed the same question a total of 20 times to the model. For the models with the higher costs per token, we reduced 10 to 5 to reduce cost. For OpenAI models, the empirical frequencies obtained closely agree with the log probabilities, but log probs were not available for all models.

Valid Inferences	**Examples**	**Controv. Conditional**	**Examples**
DS: $p \lor q, \neg q \vdash p$	Either Fido is inside or Fido is in the garden. Fido is not in the garden. $\vdash$ Fido is inside.	*CT*: $p \to q \vdash \neg q \to \neg p$	If it's raining, then it's not raining hard. $\vdash$ If it's raining hard, then it's not raining.
MP: $p \to q, p \vdash q$	If Mary was at the wedding, then Sue was at the wedding. Mary was at the wedding. $\vdash$ Sue was at the wedding.	*AS*: $p \to q \vdash (p \land r) \to q$	If the match is struck, then it will light. $\vdash$ If the match is struck and has been soaked in water, then it will light.
MT: $p \to q, \neg q \vdash \neg p$	If Mary was at the wedding, then Sue was at the wedding. Sue was not at the wedding. $\vdash$ Mary was not at the wedding.	*CMP*: $p \to (q \to r), p$	If the Warriors don't win, then if the Lakers don't win, the Celtics will.
MiN: $\Diamond\neg p \vdash \neg\Box p$	Mary might not have been at the wedding. $\vdash$ It's not the case that Mary must have been at the wedding.	$\vdash q \to r$	The Warriors won't win. $\vdash$ If the Lakers don't win, the Celtics will.
NMu: $\neg\Box p \vdash \Diamond\neg p$	It's not the case that Mary must have been at the wedding. $\vdash$ Mary might not have been at the wedding.		
Invalid Inferences		**Controv. Modal**	
AC: $p \to q, q \vdash p$	If Mary was at the wedding, then Sue was at the wedding. Sue was at the wedding. $\vdash$ Mary was at the wedding	*DSmu:* $p \lor \Box q, \neg\Box q \vdash p$	Either Fido is inside or Fido must be in the garden. It's not the case that Fido must be in the garden. $\vdash$ Fido is inside.
CONV: $p \to q \vdash q \to p$	If Mary was at the wedding, then Sue was at the wedding. $\vdash$ If Sue was at the wedding, then Mary was at the wedding.	*DSmi*: $p \lor \Box q, \Diamond\neg q \vdash p$	Either Fido is inside or Fido must be in the garden. Fido might not be in the garden. $\vdash$ Fido is inside.
DA: $p \to q, \neg p \vdash \neg q$	If Mary was at the wedding, then Sue was at the wedding. Mary was not at the wedding. $\vdash$ Sue was not at the wedding.	*MTmi*: $p \to \Box q, \Diamond\neg p \vdash \neg p$	If Mary was at the wedding, then Sue must have been there. Sue might not have been there. $\vdash$ Mary was not at the wedding.
INV: $p \to q \vdash \neg p \to \neg q$	If Mary was at the wedding, then Sue was at the wedding. $\vdash$ If Mary was not at the wedding, then Sue was not at the wedding.	*MTmu*: $p \to \Box q, \neg\Box q \vdash \neg p$	If Mary was at the wedding, then Sue must have been there. It's not the case that Sue must have been at the wedding. $\vdash$ Mary was not at the wedding.
MuDistOr: $\Box(p \lor q) \vdash \Box p \lor \Box q$	The envelope must have been upstairs or under a bed. $\vdash$ The envelope must have been upstairs or it must have been under a bed.	*WSFC*: $\Diamond p \lor \Diamond q \vdash \Diamond p \land \Diamond q$	John might clean his room or he might go to the lecture. $\vdash$ John might clean his room and he might go to the lecture.
MiAg: $\Diamond p \land \Diamond q \vdash \Diamond(p \land q)$	The envelope might be upstairs and the envelope might be under a bed. $\vdash$ The envelope might be upstairs under a bed.	*NSFC*: $\Diamond(p \lor q) \vdash \Diamond p \land \Diamond q$	John might clean his room or go to the lecture. $\vdash$ John might clean his room and John might go to the lecture.

Table 1: Key inferences tested; p, q stand for modal/conditional-free propositions, $\neg$ for 'not', $\lor$ for 'or', $\to$ for 'if...then', $\Diamond$ for 'might', and $\Box$ for 'must'. $\varphi_1, \ldots, \varphi_n \vdash \psi$ is the inference from the list of premises $\varphi_1, \ldots, \varphi_n$ to the conclusion ψ. "Controv." stands for controversial inferences. Figure 1 summarizes success on all and only the uncontroversial inferences.

overall correlation for all inferences in Table 1 is .85. While we could run more instances for all inference patterns, limited only by time and cost, the high correlations between responses to sensical and nonsensical instances suggest that we already have enough instances to see how an LLM responds to the logical form in question, not just to contingent features of particular instances of that form.

We also tested to what extent our results are sensitive to the choice of the word 'infer' in our prompts, as opposed to other phrases we consider equivalent in this context: 'deduce', 'conclude', 'logically infer', 'logically deduce', and 'logically conclude'. Could it be, e.g., that models might commit fewer conditional fallacies if we prompt them with one of these other phrases? We ran the instances of the AC inference (see Table 1) using each of the mentioned substitutes for 'infer'. The results are qualitatively the same as for 'infer', and the correlation coefficients between the yes-frequencies for 'infer' and for each of the substitutes are over .9. Thus, which of the phrases above we use apparently makes little difference.[2]

4 Results

Key inferences we tested are summarized in Table 1. To establish a baseline for performance, we tested a variety of inferences whose (in)validity is uncontroversial (as labeled in the left-hand side of the table): DS, MP, MT, AC, DA, and x, o, and ox variants thereof, as well as INV, CONV, MiN, NMu, MuDistOr, MiAg, and x variants thereof. The aggregate performance of our models on these inferences is reported in Table 2 for the temperature 0 settings, in addition to Figure 1 at the beginning of the paper. We see that the results significantly vary across models, and performance roughly correlates with model size. No model achieves above 90% accuracy on this set of tasks. Moreover, Llama 3.1 405B and 70B are the only open-weights models that achieve 80%+ accuracy, with the former being the best performer overall. The GPT-4 model family show generally strong performance.

We observe that the models perform similarly in the few-shot setting compared to the zero-shot setting (paired t-test not statistically significant). This is presumably because the few-shot examples we

[2] By contrast, there was a bigger difference when we explicitly asked LLMs, "Is this form of argument logically valid?" (we call this the 'v' variant of each question). This decreased the rate at which some LLMs accepted fallacious inferences, perhaps due to the presence of texts on formal logic in the training data. However, it also decreased the rate at which some LLMs accepted valid inferences (see the GitHub repository). In any case, the qualitative observations we make still apply with this alternative formulation of the questions.

Model $(T = 0)$	**0-shot** **Accuracy %**	**Few-shot** **Delta**	**0-shot Cot** **Delta**
Llama 3.1 Instruct 405B	89.50	0.17	1.00
GPT-4 Turbo (2024-04-09)	87.17	2.00	4.33
Claude 3.5 Sonnet	85.33	2.83	2.33
GPT-4 Turbo (1106)	84.67	1.33	1.67
GPT-4 (0613)	83.50	2.50	5.17
Llama 3.1 Instruct 70B	81.33	3.33	5.50
Gemini 1.5 Pro	81.17	-4.83	7.67
GPT-4 (0314)	80.33	5.33	7.50
GPT-4o (2024-05-13)	79.67	0.50	9.00
GPT-4o mini	75.33	8.17	13.50
Claude 3 Opus	74.00	6.00	15.83
Gemini 1.5 Flash	72.83	-2.33	13.50
Mistral Large 2	72.17	10.67	12.50
Mixtral 8x7B	71.17	-7.83	0.17
Llama 3 Instruct 70B	70.33	5.33	11.00
Claude 3 Sonnet	68.17	2.17	12.00
Claude 3 Haiku	65.50	-3.67	11.00
Mixtral 8x22B	65.17	12.50	21.50
Gemma 2 27B	64.67	1.33	16.50
Llama 3 Instruct 8B	63.83	-4.00	8.17
Code Llama 34B	58.67	-7.83	2.67
GPT-3.5 Turbo (0125)	57.33	2.33	17.00
Llama 2 Chat 7B	55.00	-1.00	3.50
GPT-3.5 Turbo (1106)	54.33	1.17	14.67
Llama 3.1 Instruct 8B	53.67	8.00	19.50
Code Llama 7B	50.33	4.50	12.67
Llama 2 Chat 13B	49.67	7.83	9.00
Mistral 7B	49.17	0.17	21.17
Code Llama 13B	45.33	-8.00	8.17

Table 2: Model performance on uncontroversial inferences with different prompting setups.

use (see Appendix B) are about conjunctions and disjunctions, not directly about conditionals and modals. The motivation of the few-shot setting is to make the logical reasoning task clear in context. The results suggest that models understand the task in the zero-shot setting, yet they are not capable of consistently recognizing valid and invalid inference patterns. On the other hand, zero-shot chain-of-thought (CoT) prompting does dramatically improve performance (paired t-test statistically significant): the best models all achieve accuracy near 90%, while still systematically making some mistakes with modal or conditionals. These results contribute to the cumulative evidence that CoT elicits and improves reasoning in LLMs (Wei et al., 2022; Kojima et al., 2022; Suzgun et al., 2022), while pointing to lacunae that persist even with CoT. The performance trends of the temperature 1 settings are similar, and we include figures for those, and for CoT prompting, in Appendix C. Next, we summarize some noteworthy findings based on specific inferences from Table 1.

4.1 Divergences from the material analysis

Under the material analysis of conditionals discussed in Section 2, CT would be valid. However, CT is not valid according to the modal analysis of conditionals mentioned in Section 2, and indeed, there are well-known intuitive counterexamples to CT (Stalnaker, 1968): for instance, the inference from 'If it's raining, it's not raining hard' to 'If it's raining hard, it's not raining', is obviously not valid, but this inference would be valid if CT were (given very weak background assumptions).

Another inference pattern that is valid according to the material analysis is AS. But there are again well-known counterexamples to AS: from 'If the match is struck, it will light', we cannot infer 'If the match is struck and it has been soaked in water, it will light' (Stalnaker, 1968).

The LLMs we tested largely agree with human judgment in rejecting CT and AS (see Appendices C.1.10-C.1.11). This underscores the importance of not assuming the material analysis when evaluating LLMs, since if we did, we would wrongly ascribe mistakes to them in this case. And it shows that LLMs, like ordinary human speakers, do not interpret the natural language conditional 'if... then' as a material conditional.

4.2 Inconsistency and overgeneralization

We tested a number of inference patterns that involve the interaction of modals with conditionals or disjunction. This is an especially interesting domain, since work in philosophy and logic has shown that substituting a modal sentence for a non-modal one can change a pattern from being apparently valid to apparently invalid.

For instance, $p \vee q, \neg q$ uncontroversially entails p (DS) *provided p and q are Boolean* (that is, do not themselves contain modals or conditionals). But it is not clear that $p \vee \Box q$, together with $\neg \Box q$, entails p (DSmu; see Klinedinst and Rothschild (2012)). For instance, if we know that Fido is either inside or outside, but don't know where he is, then it seems we know (i) Fido is either inside or else must be outside; and (ii) it's not true that Fido must be outside (since he might be inside). But we need not conclude that Fido is inside, contrary to DSmu.

Similarly, we can uncontroversially conclude $\neg p$ from $p \to q$ together with $\neg q$, when p and q are Boolean (MT). But it is not so clear that the inference from $p \to \Box q$ and $\neg \Box q$ to $\neg p$ (MTmu) is valid (Yalcin, 2012). It seems that we can know (i) if Fido is not inside, he must be outside and (ii) it's not true that Fido must be outside (since he might be inside), without being compelled to conclude that Fido is inside, contrary to MTmu.

For a final case, McGee (1985) pointed out that, while the inference from $p \to q, p$ to q is obviously valid when p, q are Boolean (MP), the inference from $p \to (q \to r), p$ to $q \to r$ (CMP) appears *invalid*. Suppose that the Lakers, Warriors, and Celtics are the only finalists in a tournament, so it's certain that (i) if the Warriors don't win, then if the Lakers don't win, the Celtics will. Suppose moreover that (ii) the Warriors are very likely not to win, because the Lakers are way ahead. But now suppose further that the Celtics are heavy underdogs, so that if the Lakers don't win, the Warriors probably will. We then can't conclude from (i) and (ii) that it's likely that, if the Lakers don't win, the Celtics will. But this would follow if CMP were valid, since valid inference preserves probability.

These cases are of special interest to test on LLMs since human subjects can immediately recognize that while DS, MT, and MP are valid for Boolean sentences, they apparently fail for modal/conditional substitution instances of these patterns like DSmu, MTmu, and CMP. That is, humans *do not overgeneralize from the simple (Boolean) case to the general case*. It is thus interesting to explore whether LLMs are human-like in this respect or rather overgeneralize from simple to complex cases—especially since these complex cases have only been discussed in a relatively small number of philosophy and logic papers and are presumably somewhat rare in naturalistic settings and hence presumably not very frequent in training data.

We found that, indeed, many LLMs do overgeneralize: they do not agree with human judgments about the invalidity of MTmu, DSmu, and CMP. Intriguingly, some models exhibit human-like judgments in rejecting MTmi, but at the same time they accept MTmu (Figure 2). This is logically inconsistent, since those models also accept that 'might not' is logically equivalent to 'not must' (MiN and NMu), [3] and MTmi and MTmu differ only by substituting these equivalent phrases. This intriguing pattern suggests that LLMs may indeed overgeneralize from the validity of MT to judging MTmu to be valid, while recognizing that MTmi (which is not syntactically an instance of MT) is invalid.[4]

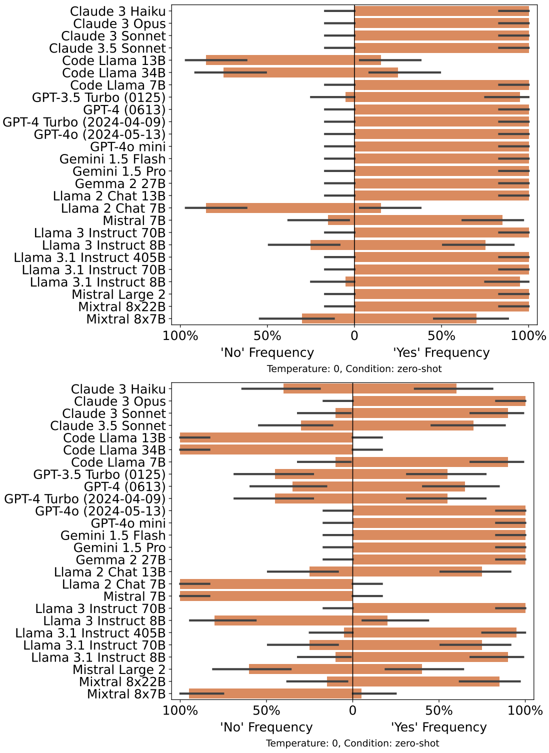


Figure 2: Zero-shot responses for MTmu (above) and MTmi (below) show inconsistency for many models. All error bars, including in subsequent figures, represent 95% confidence intervals.

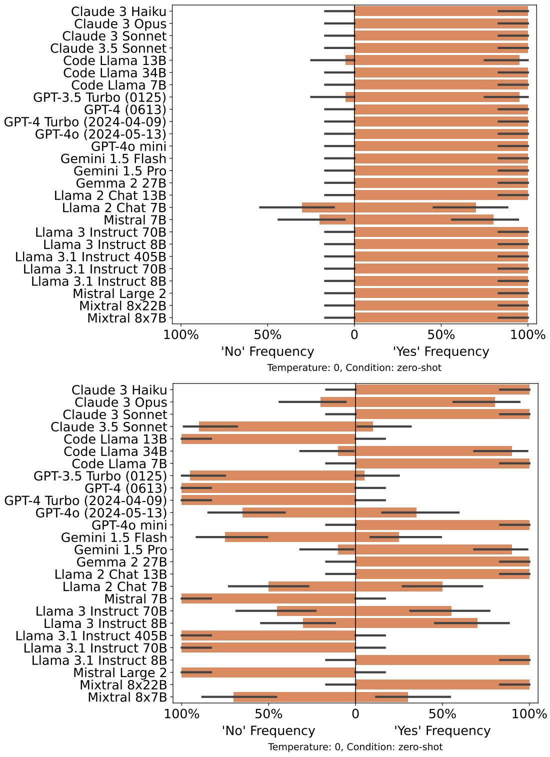


Figure 3: Zero-shot responses for DSmu (above) and DSmi (below) show inconsistency for many models.

We found similar patterns for DS, as shown in Figure 3, where many models accepted DSmu but rejected DSmi. Indeed, even the GPT-4 models we tested exhibit logically inconsistent behavior in

[3] In fact, the inconsistency arises just using MiN. For given MiN, the premises of MTmi entail the premises of MTmu, so one cannot reject the former and accept the latter.

[4] Our examples were designed to elicit an epistemic reading of the modals in question. However, this is immaterial to our logical points, since even if the target LLM instead accessed a deontic reading, the judgments we report would remain jointly inconsistent, as long as the LLM's interpretation does not change across different instances of the modal in question.

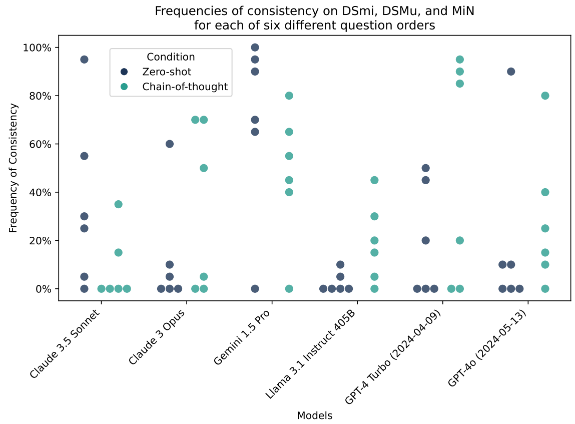


Figure 4: Percentage of responses that were jointly consistent when we asked leading models about DSmu, MiN, and DSmi in the same context window, in one of the six possible orders. Each dot represents such an order. The results show strong sensitivity to question order, which is highly undesirable.

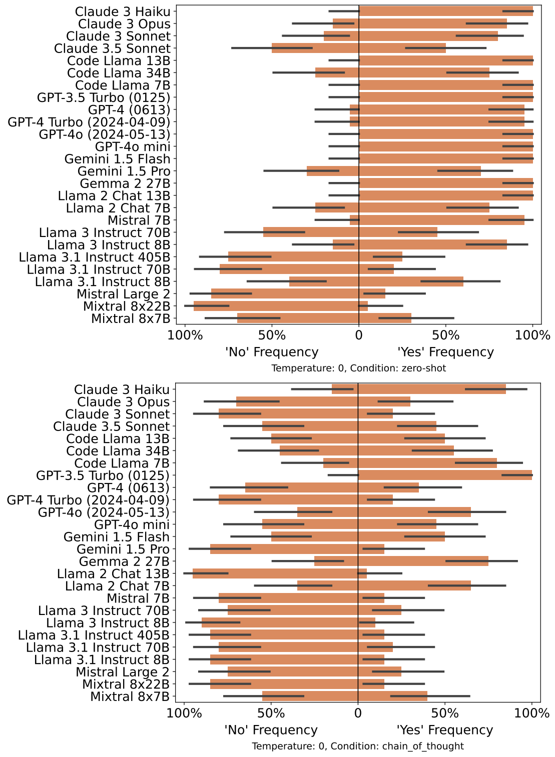


Figure 5: Responses for CMP, zero-shot (above) and chain-of-thought (below); LLMs were asked whether the inference preserved likelihood, i.e., if $q \to r$ must be likely when $p \to (q \to r)$ is certain and p is likely.

this area, accepting DSmu and MiN while rejecting DSmi. This inconsistency was displayed for both zero-shot and chain-of-thought prompts. We wondered whether this incoherence could be due to the fact that the inconsistent triad of responses was given in response to different prompts in different context windows, so we probed more deeply, asking the leading models about each of the three inference patterns within the same context window, but their responses remained logically inconsistent much of the time; see Figure 4 for a summary.

Also strikingly, the models that performed best overall on other inferences uniformly judge CMP to be valid with zero-shot prompts, contrary to human judgment. Performance improved substantially with CoT prompts, but most of the best-performing models still accept CMP at surprisingly high rates (Figure 5), pointing to a stark contrast between reflective human judgment and state of the art LLM performance. Along with our findings about MT and DS, this supports the hypothesis that LLMs may overgeneralize from the validity of inference patterns for Boolean sentences to the unrestricted validity of those patterns for all substitutions.

4.3 Modal fallacies and free choice

We also found intriguing patterns involving LLMs' behavior with purely modal inferences, shown in Appendix C.1.17-C.1.20. While, as noted, the models were generally able to correctly reason about duality (inferring 'not must' from 'might not', and vice versa), they systematically made basic errors in modal reasoning in other cases. Strikingly, almost all models, including those that performed best overall, accepted both MuDistOr and MiAg as valid, despite these being clearly invalid. (For MuDistOr, asssume the keys must be upstairs or downstairs; it doesn't follow that either they must be upstairs or they must be downstairs. After all, they might not be upstairs, and they might not be downstairs. For MiAg, assume the keys might be upstairs and might be downstairs; it doesn't follow that they might be both upstairs and downstairs.) The situation with WSFC and NSFC is also interesting. There is controversy in the semantics literature about the status of these patterns, which strike many speakers as valid but are not valid on standard modal semantics (Kamp, 1973). Intriguingly, many models accepted NSFC but rejected WSFC. In fact, this conforms to one position in the literature which maintains that, indeed, NSFC but not WSFC is valid (Simons, 2005; Meyer and Sauerland, 2017; Fusco, 2019). However, this threatens inconsistency, since most models also accept the (standardly valid) equivalence of $\Diamond(p \vee q)$ with $\Diamond p \vee \Diamond q$ (see the GitHub repository).

4.4 Relationship to some popular benchmarks

To position our logical reasoning tasks with respect to the broad landscape of LLM evaluations, we

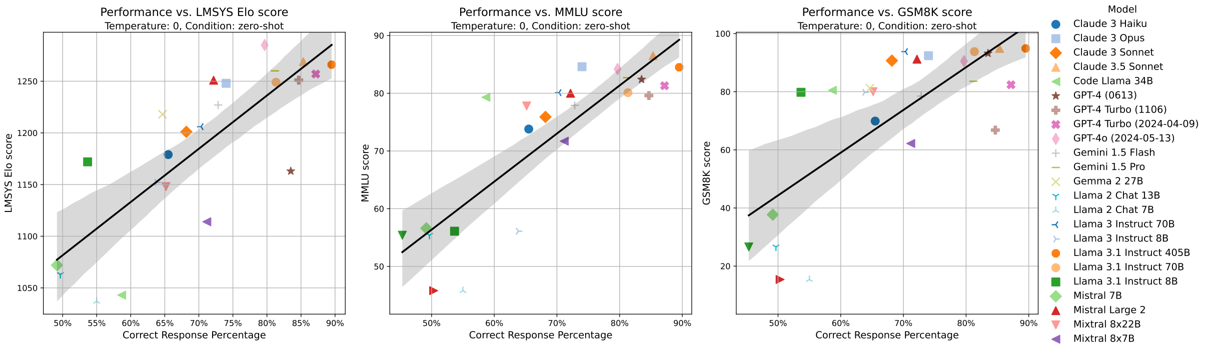


Figure 6: Correlations of our evaluation results (zero-shot) vs. LMSYS Elo ratings, MMLU scores, and GSM8k scores. The correlations are 0.81, 0.85, and 0.75, respectively. All p-values are less than 0.01.

compare the models' performance on our benchmark (uncontroversial inferences) to that of Chatbot Arena (general assistance, Chiang et al., 2024), MMLU (domain knowledge, Hendrycks et al., 2020), and GSM8K (math reasoning, Cobbe et al., 2021), all of which are popular benchmarks to assess LLM capabilities. We show that our results are highly correlated with the results of each of the three in Figure 6. The Arena Elo ratings come from LMSYS directly.[5] The MMLU and GSM8k scores are obtained from the HELM leaderboard (Liang et al., 2022).[6] The high correlations support the hypothesis that logical reasoning abilities are related to and predictive of not only mathematical reasoning abilities but also domain-general capabilities. It would be interesting to investigate the causal connections: whether improving logical reasoning also improves some general reasoning abilities.

5 Discussion

On the one hand, our results mirror what we have learned in the field over the past few years: larger models are likely to perform better at reasoning, and chain-of-thought prompting often but does not always help. On the other hand, we have identified inconsistent and counterintuitive reasoning behaviors from even the best models with and without chain-of-thought. The inference patterns that give rise to those behaviors reflect state-of-the-art research in philosophy and logic, which by their nature means they are less present in the training and fine-tuning data of LLMs (though most likely not absent). This is suggestive of sources from which we can acquire more out-of-distribution evaluation data to test LLMs, and it illustrates that LLMs' judgments may not be reliable when they encounter novel inference patterns, even if those are natural and intuitive for humans. Lastly, we note that a neurosymbolic, semantic-parsing approach to logical reasoning, as in Olausson et al. 2023, will not automatically handle the tasks we present here. To our knowledge, theorem provers do not natively implement the nuanced meaning representations we discussed in this paper, and in some cases it is still controversial what the correct semantics for certain modal operators are. Thus, working towards an automated natural language reasoning system that sensibly solves our tasks is a challenge regardless of the models and approaches.

6 Conclusion

We have drawn on philosophical, logical, and linguistic analyses of inference to explore the extent to which current LLMs are accurate logical reasoners about conditionals and modals. We hope this will be the start of a new strand of research on LLMs. There are many natural follow-ups. First, it would be interesting to compare the behavior of LLMs with experimental human subjects on all the inferences we tested. We have reported expert human judgments from the literature, but the judgments of experimental subjects might exhibit mistakes interestingly similar to or different from those we find in LLMs. Second, there are many more inference patterns to explore, involving modals, conditionals, as well as many other logical constructions, like quantifiers, attitude predicates, and degree constructions. Third, modal and conditional reasoning is also connected to probabilistic reasoning, causal models, and mental simulations, and hence could provide another perspective for studying these in LLMs and humans. More generally, we hope that there will be more work studying LLMs using philosophical, linguistic, and logical insights into the building blocks of reasoning and meaning.

[5] https://chat.lmsys.org

[6] https://crfm.stanford.edu/helm

7 Limitations

In this paper, we have focused on studying to what extent LLMs understand logical inferences with conditionals and modals. The conditional connective and 'must' and 'might' operators are the target, while we do not systematically study other important logical operators such as 'and' and 'or', or quantifiers such as 'some' and 'all'. Even in the modal domain, operators like 'probably' and 'certainly' also deserve careful analysis. In particular, the interactions of many of these different kinds of operators could lead to interesting inference patterns, which would be worthy of future study. Additionally, we employ simple syntactical constructions to create the evaluation data. While this is natural for testing logical inference, the models' performance on an inference pattern may not generalize when it is made of more complex phrases. Lastly, our dataset is in English, and the models' logical inference abilities may differ by language. We hope future work could study the multilingual aspect of logical inference as well.

Acknowledgments

For helpful feedback, we thank Dominic Hughes, James Kirkpatrick, Roger Levy, Hailey Schoelkopf, Shane Steinert-Threlkeld, and the reviewers for EMNLP 2024.

References

Ari Beller and Tobias Gerstenberg. 2023. A counterfactual simulation model of causal language. https://doi.org/10.31234/osf.io/xv8hf.

Johan van Benthem. 2008. A brief history of natural logic.

Johan van Benthem. 2023. The logic of conditionals on outback trails. *Logic Journal of the IGPL*, 31(6):1135–1152.

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2023. Learning to teach large language models logical reasoning. *arXiv preprint arXiv:2310.09158*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Igor Douven and Sara Verbrugge. 2013. The probabilities of conditionals revisited. *Cognitive Science*, 37(4):711–730.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dorothy Edgington. 1995. On conditionals. *Mind*, 104(414):235–329.

Paul Egré and Hans Rott. 2021. The Logic of Conditionals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. Metaphysics Research Lab, Stanford University.

Jonathan St BT Evans and David E Over. 2004. *If: Supposition, pragmatics, and dual processes*. Oxford University Press.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Y Vardi. 1995. *Reasoning about knowledge*. MIT Press, Cambridge, Mass.

Nir Friedman and Joseph Y Halpern. 1994. On the complexity of conditional logics. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference (KR'94)*, pages 202–213, San Francisco, CA. Morgan Kaufmann.

Melissa Fusco. 2019. Sluicing on free choice. *Semantics and Pragmatics*, 12(20):1–22.

James Garson. 2024. Modal logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy, Spring 2024 Edition*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Allan Gibbard and William L Harper. 1981. Counterfactuals and two kinds of expected utility. In William L. Harper, Robert Stalnaker, and Glenn Pearce, editors, *Ifs: Conditionals, Beliefs, Decision, Chance, and Time*, pages 153–192. D. Reidel Publishing Company.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

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.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hans Kamp. 1973. Free choice permission. In *Proceedings of the Aristotelian Society*, volume 74, pages 57–74.

Jerrold J Katz. 1972. *Semantic theory*. Harper & Row.

Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. LAMBADA: Backward chaining for automated reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568, Toronto, Canada. Association for Computational Linguistics.

Justin Khoo. 2022. *The Meaning of* If. Oxford University Press.

Nathan Klinedinst and Daniel Rothschild. 2012. Connectives without truth-tables. *Natural Language Semantics*, 20:137–175.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Angelika Kratzer. 1981. The notional category of modality. In H. Eikmeyer and H. Rieser, editors, *Words, Worlds, and Contexts: New Approaches in Word Semantics*, pages 38–74. de Gruyter.

Angelika Kratzer. 2012. *Modals and Conditionals*. Oxford University Press.

Saul A Kripke. 1963. Semantical analysis of modal logic I. Normal modal propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9:67–96.

David Lewis. 1973a. Causation. *The Journal of Philosophy*, 70(17):556–567.

David Lewis. 1973b. *Counterfactuals*. Basil Blackwell, Oxford.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the eight international conference on computational semantics*, pages 140–156.

Vann McGee. 1985. A counterexample to modus ponens. *The Journal of Philosophy*, 82(9):462–471.

Marie-Christine Meyer and Uli Sauerland. 2017. Covert across-the-board movement revisited: Free choice and the scope of modals. In *Proceedings of NELS 47*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

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

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah D Goodman. 2023. Certified deductive reasoning with language models.

Paul Portner. 2009. *Modality*. Oxford University Press.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

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

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using ood examples. In *Advances in Neural Information Processing Systems*, volume 36, pages 3083–3105. Curran Associates, Inc.

Mandy Simons. 2005. Dividing things up: The semantics of *or* and the modal/*or* interaction. *Natural Language Semantics*, 13(3):271–316s.

Robert C Stalnaker. 1968. A theory of conditionals. In Nicholas Rescher, editor, *Studies in Logical Theory*, pages 98–112. Blackwell.

Robert C Stalnaker and Richmond H Thomason. 1970. A semantic analysis of conditional logic. *Theoria*, 36(1):23–42.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2020. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*.

Alfred Tarski. 1936. Über den Begriff der logischen Folgerung. *Aces du Congrès International de Philosophie Scientifique*, fasc. 7:1–11.

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

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

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael R Lyu. 2024. A & b== b & a: Triggering logical reasoning failures in large language models. *arXiv preprint arXiv:2401.00757*.

Albert Webson, Alyssa Marie Loo, Qinan Yu, and Ellie Pavlick. 2023. Are language models worse than humans at following prompts? it's complicated. *arXiv preprint arXiv:2301.07085*.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*.

Seth Yalcin. 2012. A counterexample to modus tollens. *Journal of Philosophical Logic*, 41:1001–1024.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023. Satlm: Satisfiability-aided language models using declarative prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 45548–45580. Curran Associates, Inc.

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

A Language models used

In this section, we cite all the LLMs used to conduct our experiments: the GPT-4 model family (OpenAI, 2023); the GPT-3.5 model family (Brown et al., 2020);[7] the Claude 3 model family[8] and Claude 3.5 Sonnet;[9] Gemini 1.5 Pro and Flash (Reid et al., 2024); Gemma 2 27B (Gemma Team, 2024); the Llama 3 model family (Dubey et al., 2024); Mistral Large 2;[10] Mixtral 8x7B (Jiang et al., 2024) and Mistral 7B (Jiang et al., 2023); Llama 2 Chat 7B and 13B (Touvron et al., 2023); Code Llama 7B, 13B, and 34B (Roziere et al., 2023).

B Prompts used in the experiments

We use the following prompts to run the experiments reported in this work. The prompts are the same for all models, except that for some small models (e.g., Llama 2 7B and Code Llama 7B), we had to modify the chain-of-thought prompt with extra reminders to answer only after first thinking step by step. Each prompt is shown with a concrete example of an inference pattern.

B.1 Zero-shot

```
Answer only with 'yes' or 'no' and nothing else.

From 'If Mary was at the wedding, then Sue was
    ↪ at the wedding' together with 'Sue was
    ↪ not at the wedding', can we infer 'Mary
    ↪ was not at the wedding'?
```

B.2 Few-shot

```
Consider the following examples:

From 'Ann went to the store', can we infer that
    ↪ 'Ann went to the store and Bob went to
    ↪ the beach'? Correct answer: No.

From 'Ann went to the store', can we infer that
    ↪ 'Ann went to the store or Bob went to the
    ↪  beach'? Correct answer: Yes.

From 'Ann went to the store and Bob went to the
    ↪ beach', can we infer that 'Ann went to
    ↪ the store'? Correct answer: Yes.

From 'Ann went to the store or Bob went to the
    ↪ beach', can we infer that 'Ann went to
    ↪ the store'? Correct answer: No.

Now here is a question for you:

From 'If Mary was at the wedding, then Sue was
    ↪ at the wedding' together with 'Sue was
    ↪ not at the wedding', can we infer 'Mary
    ↪ was not at the wedding'?
```

B.3 Zero-shot chain-of-thought

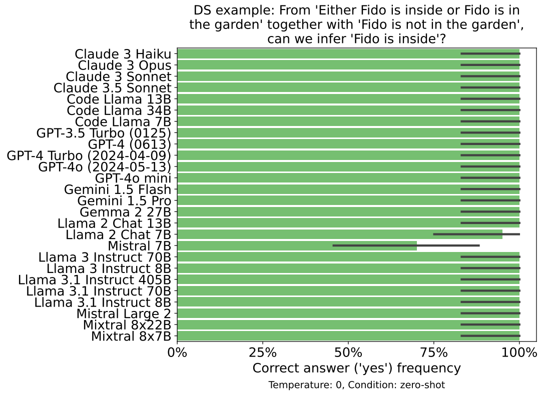
```
In response to the following question, think
    ↪ step by step and explain your reasoning;
    ↪ then when you are ready to answer, simply
    ↪  write 'Answer: ' followed by 'yes' or '
    ↪ no'.

From 'If Mary was at the wedding, then Sue was
    ↪ at the wedding' together with 'Sue was
    ↪ not at the wedding', can we infer 'Mary
    ↪ was not at the wedding'?
```

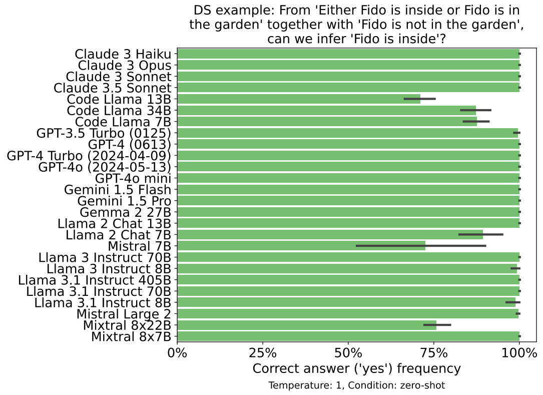
C Additional results

C.1 Individual inferences

In this Appendix, we display the zero-shot performance of the LLMs on the inferences shown in Table 1 when operating at both temperature 0 and temperature 1.

C.1.1 Disjunctive Syllogism (DS)





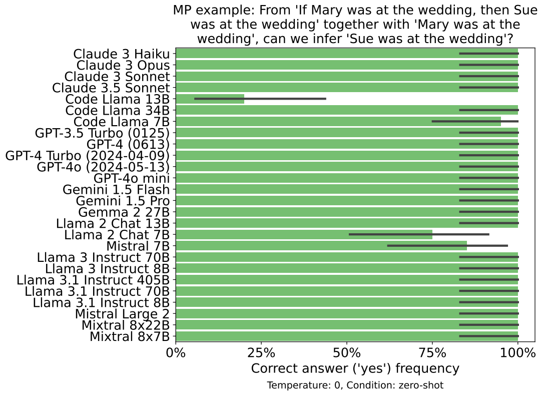
[7] https://openai.com/index/chatgpt

[8] https://www.anthropic.com/news/claude-3-family

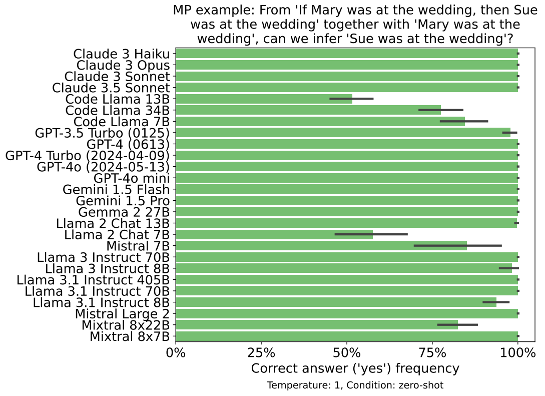
[9] https://www.anthropic.com/news/claude-3-5-sonnet

[10] https://mistral.ai/news/mistral-large-2407

C.1.2 Modus Ponens (MP)

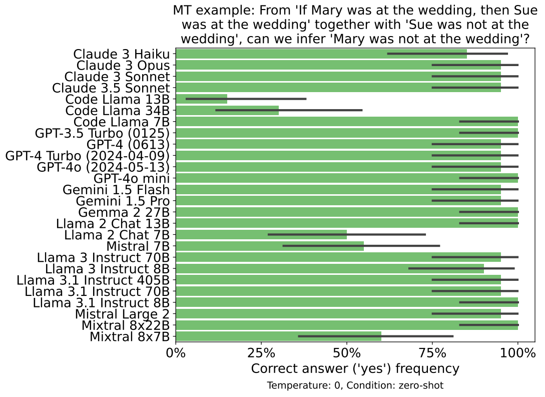

MP example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Mary was at the wedding', can we infer 'Sue was at the wedding'?

Temperature: 0, Condition: zero-shot

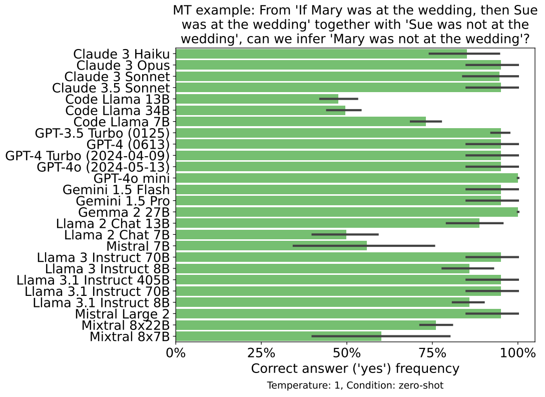

MP example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Mary was at the wedding', can we infer 'Sue was at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.3 Modus Tollens (MT)

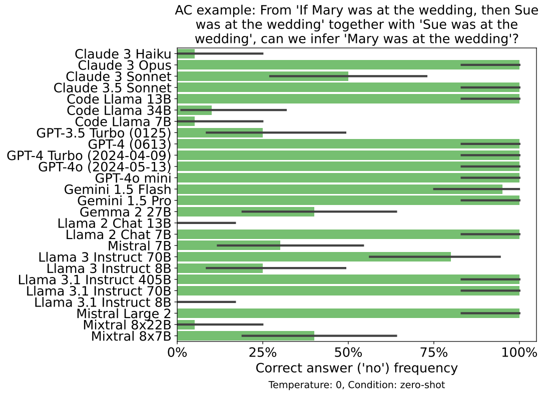

MT example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Sue was not at the wedding', can we infer 'Mary was not at the wedding'?

Temperature: 0, Condition: zero-shot

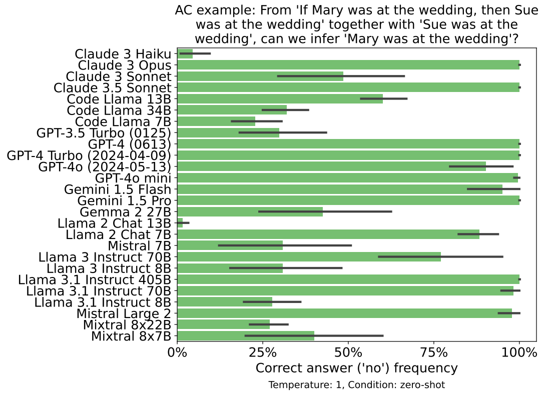

MT example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Sue was not at the wedding', can we infer 'Mary was not at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.4 Affirming the Consequent (AC)

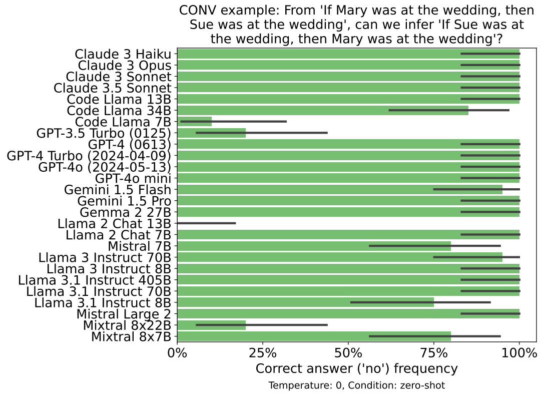

AC example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Sue was at the wedding', can we infer 'Mary was at the wedding'?

Temperature: 0, Condition: zero-shot

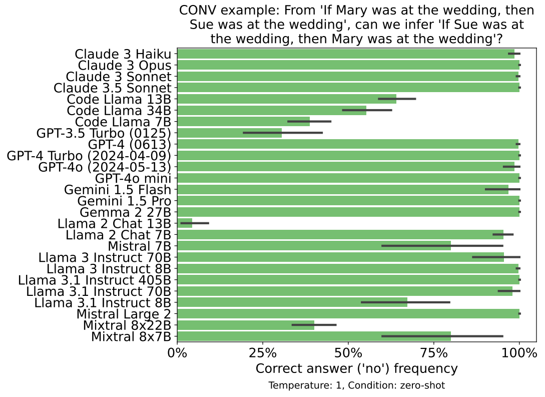

AC example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Sue was at the wedding', can we infer 'Mary was at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.5 Conversion CONV)

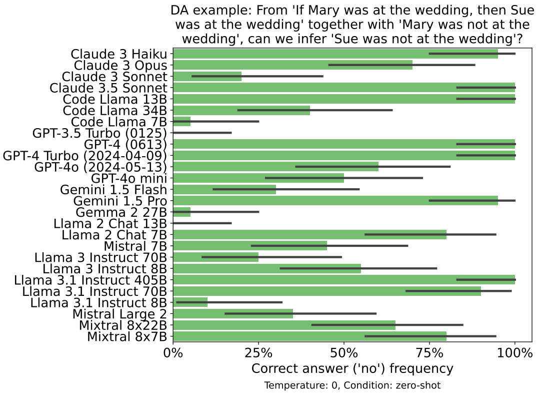

CONV example: From 'If Mary was at the wedding, then Sue was at the wedding', can we infer 'If Sue was at the wedding, then Mary was at the wedding'?

Temperature: 0, Condition: zero-shot

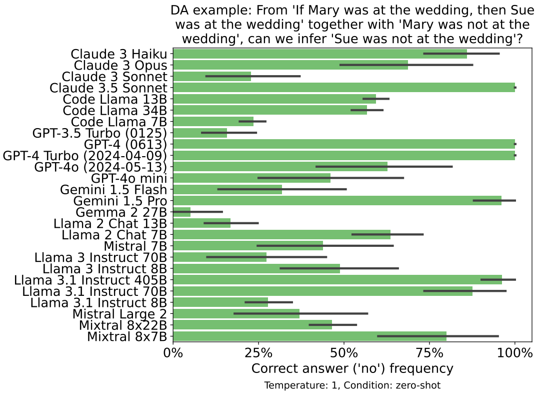

CONV example: From 'If Mary was at the wedding, then Sue was at the wedding', can we infer 'If Sue was at the wedding, then Mary was at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.6 Denying the Antecedent (DA)

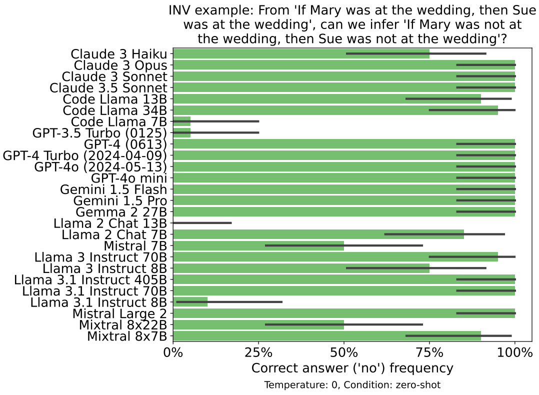

DA example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Mary was not at the wedding', can we infer 'Sue was not at the wedding'?

Temperature: 0, Condition: zero-shot

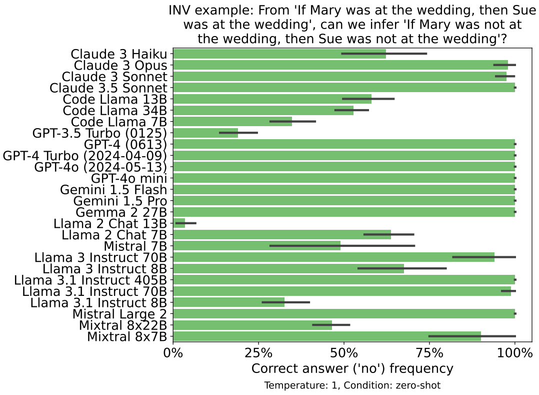

DA example: From 'If Mary was at the wedding, then Sue was at the wedding' together with 'Mary was not at the wedding', can we infer 'Sue was not at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.7 Inversion (INV)

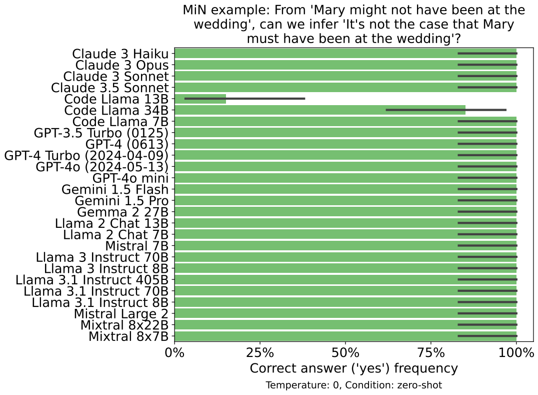

INV example: From 'If Mary was at the wedding, then Sue was at the wedding', can we infer 'If Mary was not at the wedding, then Sue was not at the wedding'?

Temperature: 0, Condition: zero-shot

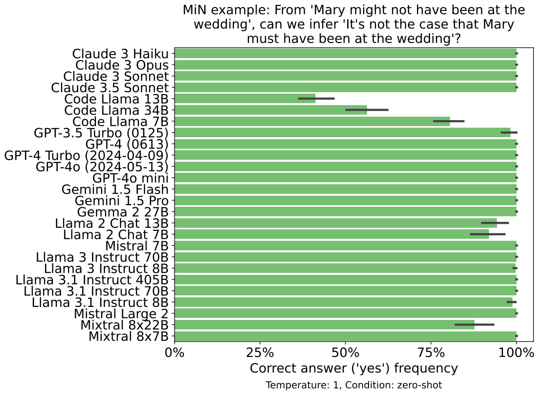

INV example: From 'If Mary was at the wedding, then Sue was at the wedding', can we infer 'If Mary was not at the wedding, then Sue was not at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.8 Might Not (MiN)

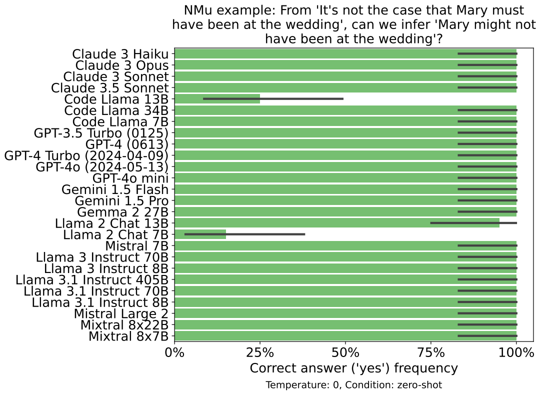

MiN example: From 'Mary might not have been at the wedding', can we infer 'It's not the case that Mary must have been at the wedding'?

Temperature: 0, Condition: zero-shot

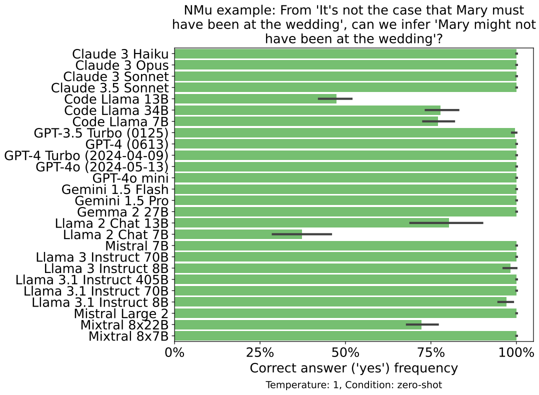

MiN example: From 'Mary might not have been at the wedding', can we infer 'It's not the case that Mary must have been at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.9 Not Must (NMu)


NMu example: From 'It's not the case that Mary must have been at the wedding', can we infer 'Mary might not have been at the wedding'?

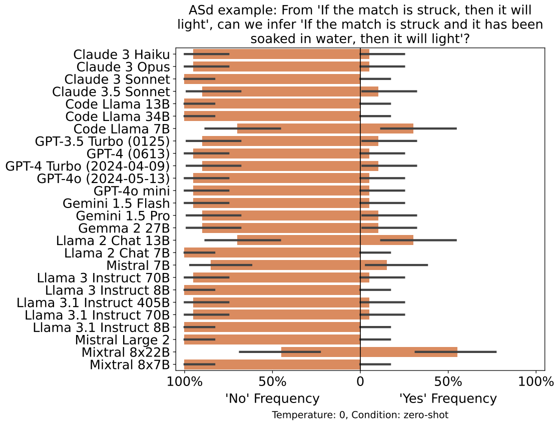
Temperature: 0, Condition: zero-shot

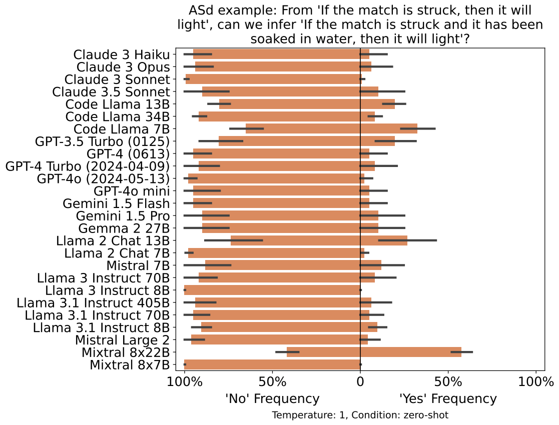

NMu example: From 'It's not the case that Mary must have been at the wedding', can we infer 'Mary might not have been at the wedding'?

Temperature: 1, Condition: zero-shot

C.1.10 Antecedent Strengthening (AS)

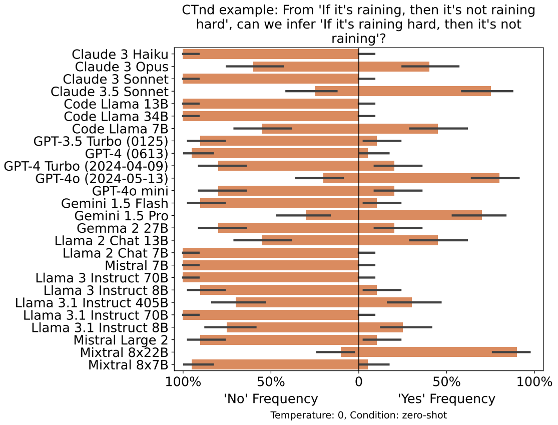
In the graphs below, the ‘d’ in ‘ASd’ indicates *deviant* instances, designed to bring out the invalidity of antecedent strengthening as in Stalnaker 1968.

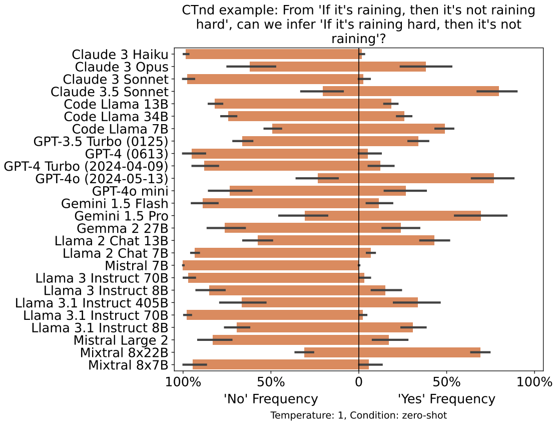




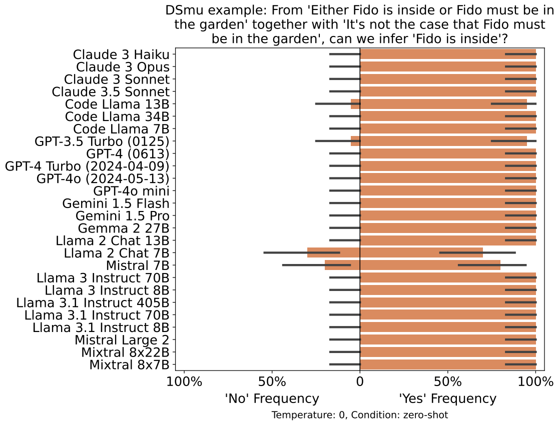
C.1.11 Contraposition (CT)

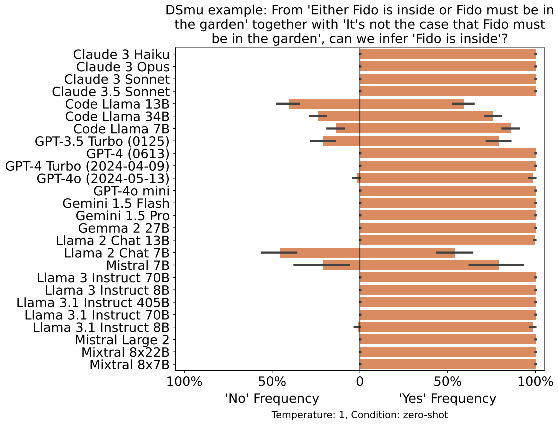
In the graphs below, the ‘n’ in ‘CTnd’ indicates that we use a version of contraposition with negation in the premise ($p \to \neg q \vdash q \to \neg p$), while the ‘d’ again indicates *deviant* instances, designed to bring out the invalidity of CT as in Stalnaker 1968.



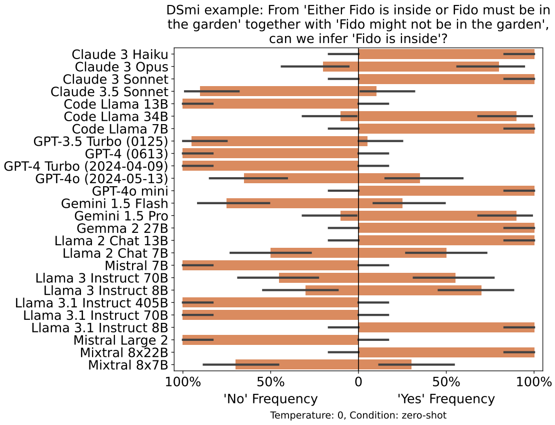


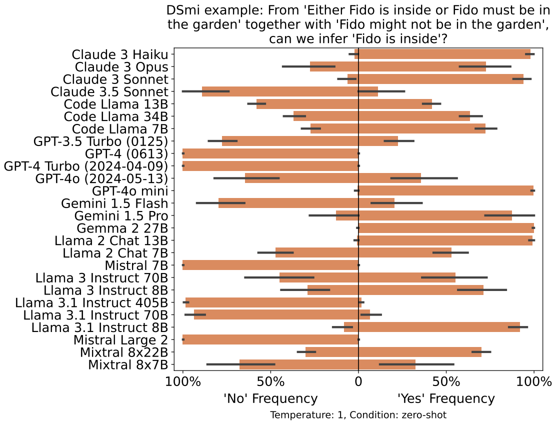
C.1.12 DS with ‘must’ (DSmu)



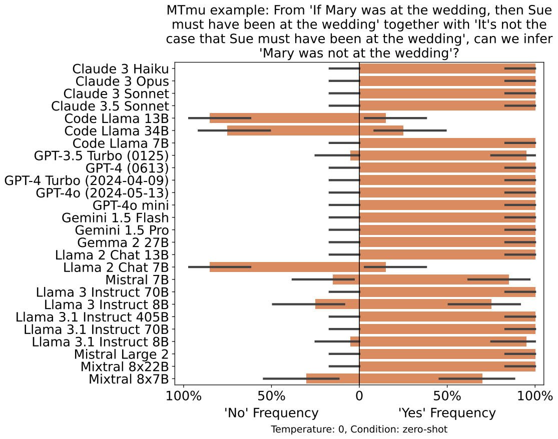


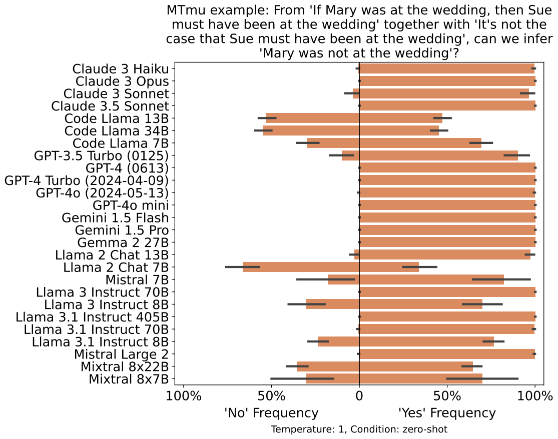
C.1.13 DS with ‘might’ (DSmi)



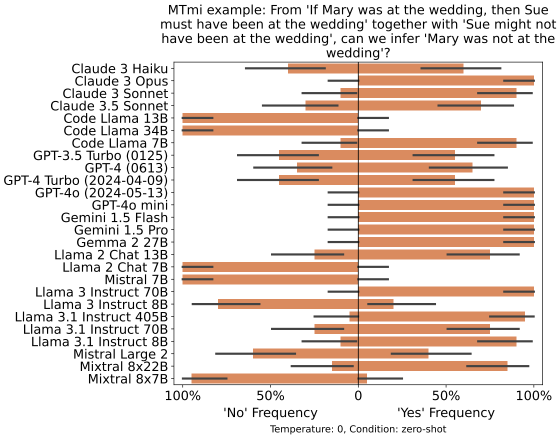


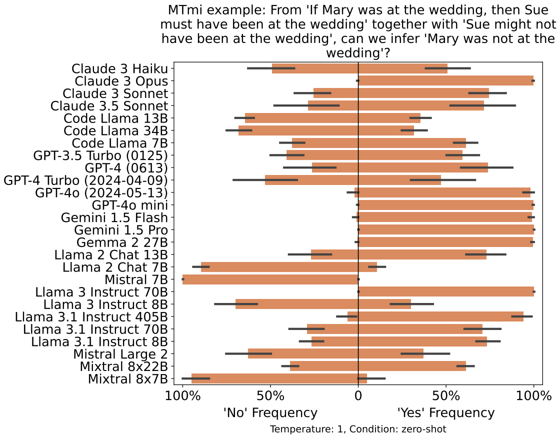
C.1.14 MT with ‘must’ (MTmu)



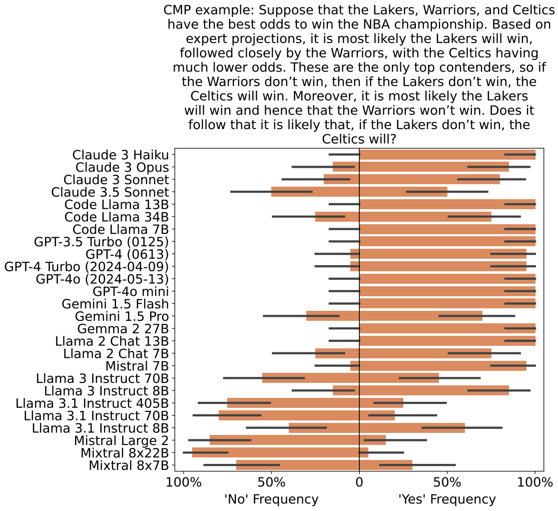


C.1.15 MT with ‘might’ (MTmi)

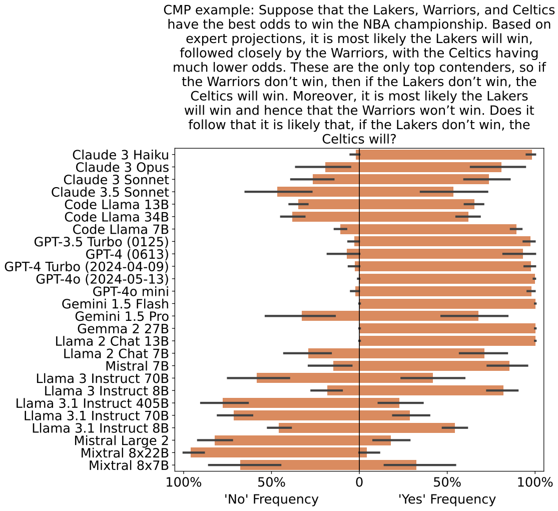




C.1.16 Complex Modus Ponens (CMP)

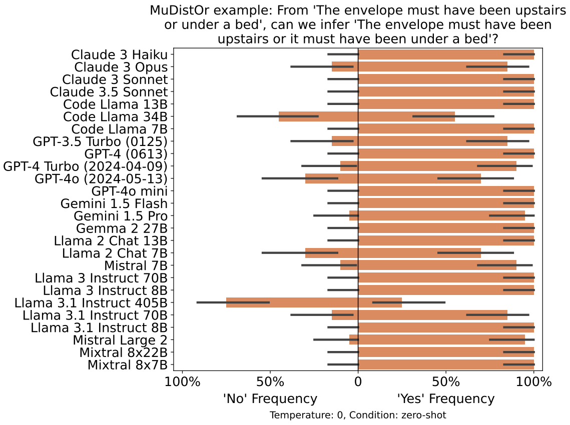


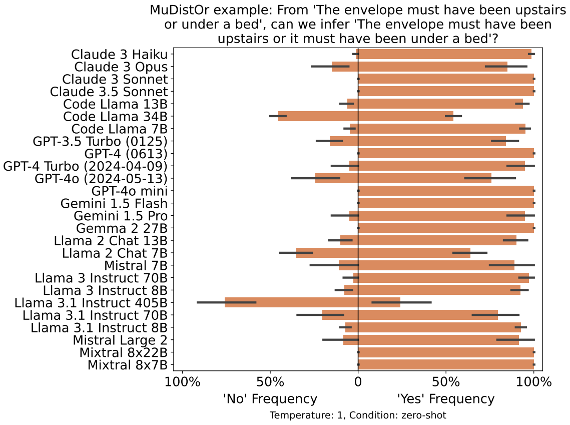
Temperature: 0, Condition: zero-shot



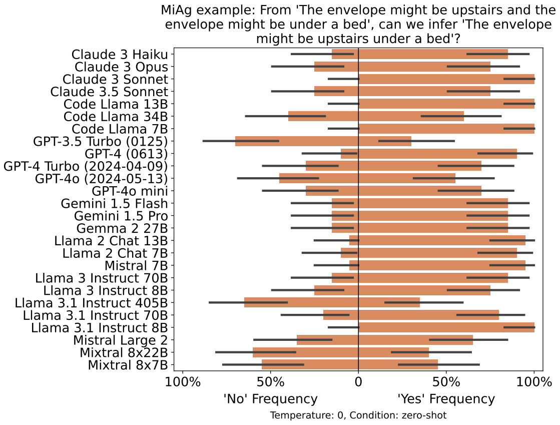
Temperature: 1, Condition: zero-shot

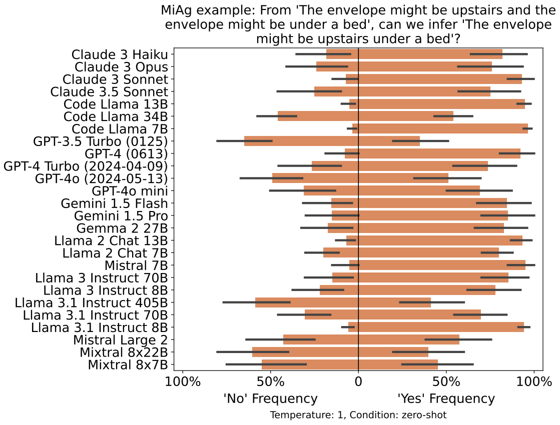
C.1.17 ‘Must’ distribution over ‘or’ (MuDistOr)



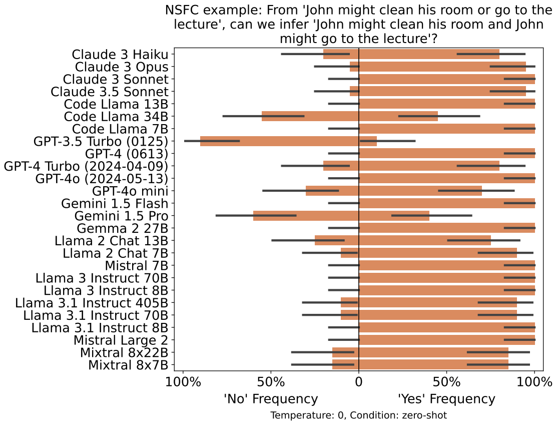


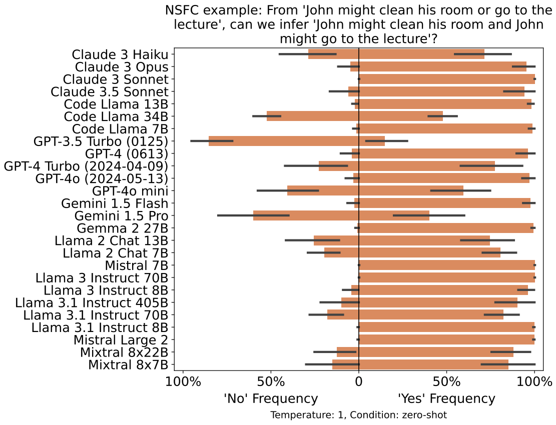
C.1.18 ‘Might’ agglomeration (MiAg)



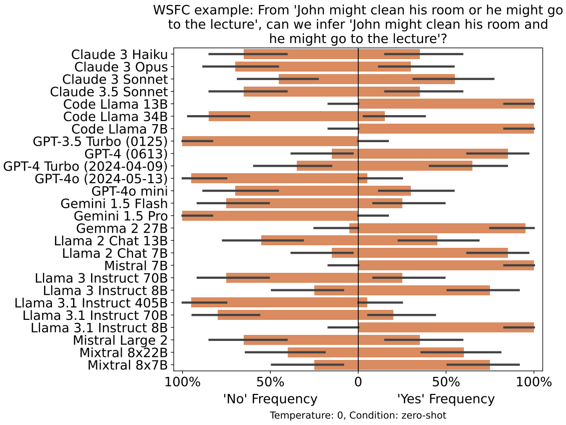


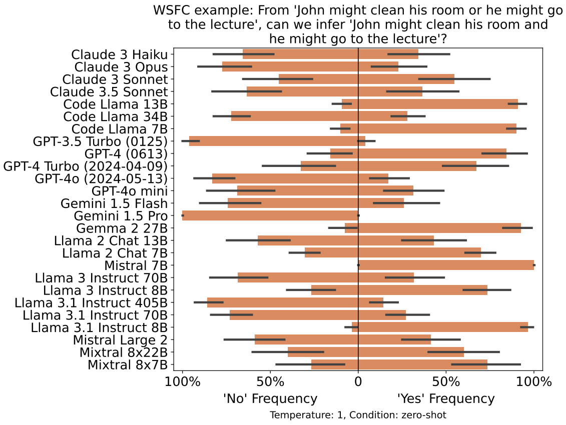
C.1.19 Narrow-scope free choice (NSFC)





C.1.20 Wide-scope free choice (WSFC)





C.2 Performance summaries

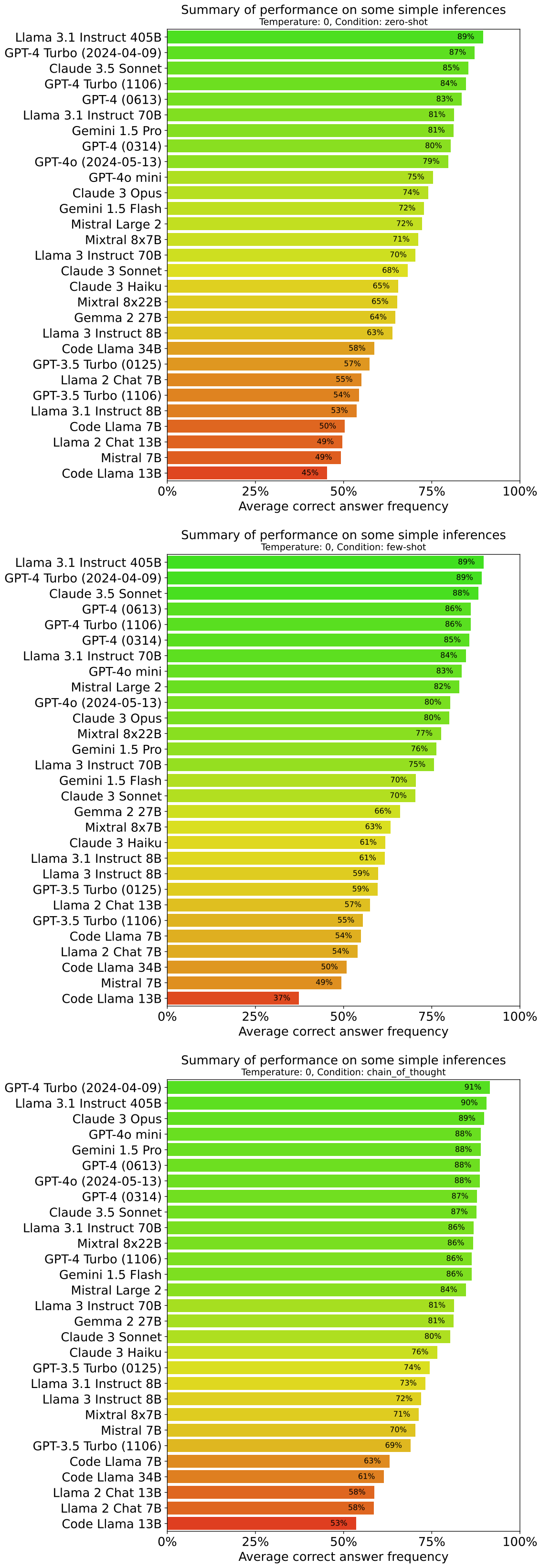


Figure 7: Summary of performance on the uncontroversial logical inference patterns under different conditions and temperature 0.

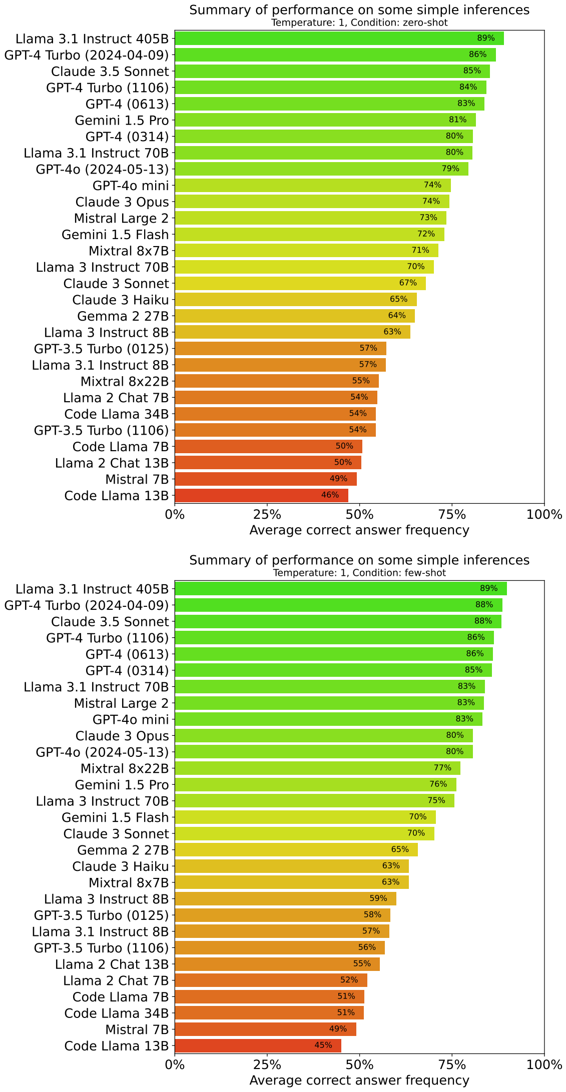


Figure 8: Summary of performance on the uncontroversial logical inference patterns under zero-shot and few-shot conditions with temperature 1.