Logical forms complement probability in understanding language model (and human) performance

Yixuan Wang
University of Chicago
yixuanwang@uchicago.edu

Freda Shi
University of Waterloo
Vector Institute, Canada CIFAR AI Chair
fhs@uwaterloo.ca

Abstract

With the increasing interest in using large language models (LLMs) for planning in natural language, understanding their behaviors becomes an important research question. This work conducts a systematic investigation of LLMs' ability to perform logical reasoning in *natural language*. We introduce a controlled dataset of hypothetical and disjunctive syllogisms in propositional and modal logic and use it as the testbed for understanding LLM performance. Our results lead to novel insights in predicting LLM behaviors: in addition to the probability of input (Gonen et al., 2023; McCoy et al., 2024), logical forms should be considered as important factors. In addition, we show similarities and discrepancies between the logical reasoning performances of humans and LLMs by collecting and comparing behavioral data from both.

1 Introduction

Logical reasoning is a fundamental aspect of building AI systems for reliable decision-making (Kautz et al., 1992, *inter alia*)—given a set of premises, an AI system should be able to deduce valid conclusions. With the advent of large language models (LLMs; Touvron et al., 2023; Jiang et al., 2023; AI@Meta, 2024, *inter alia*), there has been a surge of interest in using these models to assist planning and decision-making (Huang et al., 2022, *inter alia*); therefore, understanding the logical reasoning capabilities becomes crucial in understanding the reliability and potential of LLMs in planning. While recent work has shown that LLMs exhibit decent performance on logical reasoning problems (Liu et al., 2020; Ontanon et al., 2022; Wan et al., 2024, *inter alia*), there is still a lack of fine-grained understanding of the logical forms—among many argument forms presented in natural language (Shieber, 1993), do LLMs perform equally well, or do they exhibit preferences for certain argument forms? Do more complex components of logical forms, such as modalities, matter for LLM performance?

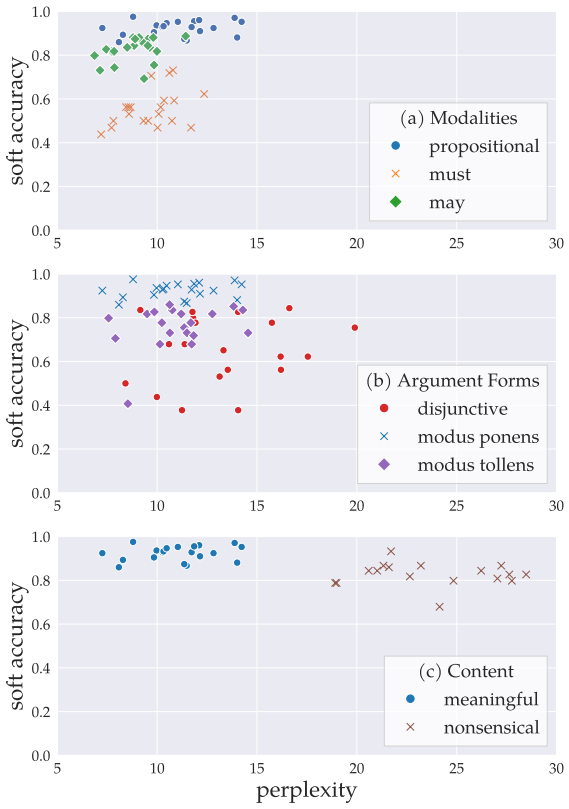

Figure 1: Illustration of the fact that perplexity does not serve as a reliable indicator of logical reasoning performance; and therefore, neither does probability. The distributions of the probabilities assigned to the ground-truth answer (i.e., soft accuracy; Y-axis) by Llama-3-70B are plotted against the perplexity of the corresponding example question (X-axis) and grouped by (a) modality, (b) argument forms, and (c) logic interpretation content. Each group consists of 20 randomly selected examples with other factors controlled.

In this work, we investigate the logical reasoning capabilities of LLMs by assessing their performance on different logical forms. We curate a dataset of natural language statements and questions based on several logical forms in both proposi-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16862–16877
July 27 - August 1, 2025

tional and modal logic, which is designed to mirror reasoning in daily communication. An example is shown in §3.3. We then conduct a series of controlled experiments to analyze the performance of a set of LLMs on the dataset. Although our findings generally align with those by Gonen et al. (2023) and McCoy et al. (2024), who suggest that LLMs excel on examples with high probability, our results indicate that logical form, including but not limited to modalities and argument forms, is a crucial complementary factor in predicting the performance of LLMs (Figure 1). Additionally, with meaningful real-world interpretations, we find that:

1. LLMs are still far from being perfect in atomic-level propositional and modal logic reasoning.
2. LLMs prefer an affirmative answer under the modality of possibility, whereas they prefer a negative answer under the modality of necessity.
3. In line with the recent results on categorical syllogisms (Eisape et al., 2024), we verify on hypothetical and disjunctive syllogisms that LLMs achieve better performance on certain logical forms that humans perform well. However, some logical forms receive favor from LLMs, while the phenomena lack support from human intuition or human behavioral data.

This paper is structured as follows. After reviewing related work (§2), we describe the dataset synthesis process (§3). We report the LLM reasoning results on our data (§4) and compare them with human performance (§5). We conclude by discussing the implications of our results and the limitations(§6).

2 Related Work

Logical reasoning benchmarks. Existing LLM logical reasoning benchmarks (Liu et al., 2020; Han et al., 2024, *inter alia*) focus on complex, multi-hop reasoning problems with manually annotated problems, making cross-problem comparisons challenging. Recent work has introduced benchmarks with synthesized natural-language questions using predefined logical formulas and substitution rules (Saparov and He, 2022; Saparov et al., 2023; Parmar et al., 2024; Wan et al., 2024, *inter alia*). Compared to them, our work uniquely incorporates modal logic, which has been largely unexplored in existing benchmarks—while Holliday et al. (2024) present a case study, our approach offers two key advances: controlled knowledge bias in logic interpretations (§3.3) and a more rigorous statistical evaluation framework (§4.1).

Propositional and modal logic reasoning in language models. Recent work has explored training and fine-tuning language models specifically for logical reasoning (Clark et al., 2021; Hahn et al., 2021; Tafjord et al., 2022). Our work differs in two key aspects: (1) we evaluate general-purpose language models through prompting, a cost-efficient setup that has been widely adopted in recent years, and we focus on propositional and alethic modal logic rather than temporal (Hahn et al., 2021) or epistemic (Sileo and Lernould, 2023) logic; [1] (2) unlike studies comparing LLM and human performance on categorical syllogisms (Eisape et al., 2024, *inter alia*),[2] we focus on hypothetical and disjunctive syllogisms with considerations of modality.

Human logic reasoning and reasoning bias. Work on human reasoning capabilities has informed studies of LLM logical reasoning: Eisape et al. (2024) compared LLM syllogistic reasoning with human behavior results (Ragni et al., 2019) under the framework of the Mental Models Theory (Johnson-Laird, 1983); Lampinen et al. (2024) found similar content effects in human and LLM reasoning, supporting the need to control for common-sense knowledge in benchmarks (§3.2); Belem et al. (2024) studied human and LLM perception of uncertainty at a lexical level. Seals and Shalin (2024) simulated the famous Wason selection task (Wason, 1968) and covered content effects over conditional syllogisms. Ozeki et al. (2024) investigated the human-like bias of LLM first-order logic reasoning introduced by content effects or answer argument forms.

Compared to them, we focus on the bias introduced by argument forms within the propositional and modal logic reasoning process. We also contribute new human behavioral data beyond the LLM results.

3 Dataset

We curate a dataset of natural-language multi-choice questions to measure the logical inference performance of LLMs. Starting from propositional and modal logical forms as templates (§3.1), we as-

[1] Technically, any logic that involves non-truth-functional operators, including first-order logic, temporal logic, and epistemic logic, can be viewed as a modal logic; however, we adopt the most restrictive sense of *modal logic* (Ballarin, 2023) and use it interchangeably with *alethic modal logic*.

[2] We refer readers to Zong and Lin (2024) for a more comprehensive review of categorical syllogisms.

sign meanings (e.g., real-world interpretations) to each variable and translate templates into natural-language Yes/No questions (§3.2). A subsidiary visualization of the process is shown in Figure 2.

3.1 Background: Propositional and Modal Logic

Propositional logic studies the relation between propositions. In this framework, each proposition is typically represented by a variable, and multiple propositions combine with logical connectives (e.g., $\vee$ and $\rightarrow$) to form compound propositions.

In propositional logic, a proposition can be evaluated as either true or false; however, this system can be overly simplistic when dealing with the complexity of real-world events. Consider the statement *Alice is not eating*, while it is true in a world where Alice is not eating, it may become false in a hypothetical *possible world* where Alice is indeed eating. This idea, known as possible world semantics (Kripke, 1959), provides a framework for more nuanced statements about event possibilities, such as *Alice may be eating* and *Alice must be eating*. The former statement can be understood as there *exists* a possible world where Alice is eating, and the latter can be understood as in *all* possible worlds, Alice is eating.[3] Normal modal logic (Kripke, 1963) formalizes this idea and extends propositional logic to reason about event necessity and possibility. In the Backus–Naur form, a normal modal logic system $\mathcal{L}$ can be written as

$$\begin{aligned}\mathcal{L} : \varphi \coloneqq p \mid \neg\varphi \mid \Box\varphi \mid \Diamond\varphi \mid \\ \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \varphi \rightarrow \varphi,\end{aligned} \tag{1}$$

where p is a propositional variable that serves as an atom in $\mathcal{L}$, $\neg$ is the negation operator, $\Box$ is the necessity operator (*must*), $\Diamond$ is the possibility operator (*may*), $\vee$ is logical disjunction (*or*), $\wedge$ is logical conjunction (*and*), and $\rightarrow$ is the logical implication operator (*if...then*). φ denotes the syntactic category of a formula in $\mathcal{L}$. The right-hand side of Eq. (1) describes all possible logical formulas under the system $\mathcal{L}$: for example, if $\varphi \in \mathcal{L}$, the rules imply that $\neg\varphi \in \mathcal{L}$, $\Box\varphi \in \mathcal{L}$, and so on. Following the convention in logic, the operator precedence is $\{\neg, \Box, \Diamond\} \succ \{\vee, \wedge\} \succ \{\rightarrow\}$.

Indeed, the operators $(\neg, \Box, \rightarrow)$ forms a functional complete set of operators under $\mathcal{L}$. Suppose φ and ψ are variables that represent logical formulas. The logical or ($\vee$) and logical and ($\wedge$) operators can be rewritten with logical not ($\neg$) and logical implication ($\rightarrow$), as follows:

$$\begin{aligned}\varphi \vee \psi &\Leftrightarrow \neg\varphi \rightarrow \psi, \\ \varphi \wedge \psi &\Leftrightarrow \neg(\varphi \rightarrow \neg\psi).\end{aligned} \tag{2}$$

Possibility operator $\Diamond$ can also be derived from the necessity operator.

$$\Diamond\varphi \Leftrightarrow \neg\Box\neg\varphi \tag{3}$$

Deduction and sequent. Given a formula set Γ as premises, if a deduction to a conclusion φ exists using axiom schemata and inference rules under the normal modal logic, we say the premises *infer* the conclusion, and the deduction can be represented as a logic *sequent* $\Gamma \vdash \varphi$. If a formula set Γ do not infer the conclusion, we denote it as $\Gamma \nvdash \varphi$ and call it a *non-entailment*.

3.2 Translating Logic to Natural Language

An *interpretation* maps propositional variables to concrete meanings. For example, under the interpretation that p is "*Jane is eating apples*" and q is "*John is eating oranges*", the logical formula $p \vee q$ becomes "*Jane is eating apples or John is eating oranges*."

Choices of interpretation, i.e., the concrete content of the sentence, should not affect the underlying logical reasoning process. However, in natural-language utterances, reasoning can be influenced by various confounding factors. Knowledge bias is a common pitfall. For example, given the logical form $\{\Box p \rightarrow \Box\neg q, \Box p\} \vdash \Box\neg q$, regardless of p's interpretation, if we interpret $\neg q \coloneqq$ *"Cats are not animals"* then the conclusion will be "*It is certain that cats are not animals.*" But common-sense knowledge suggests that "*it is certain that cats are animals*" ($\Box q$), which logically contradicts the existing premise set.[4] Such bias will complicate logical reasoning (Lampinen et al., 2024) and should be avoided in data curation. Besides, each variable should have independent interpretation, as detailed in Appendix B.1.

After being assigned interpretations, each logical form is further articulated as a yes-no question on whether the conclusion can be inferred from the

[3] The possible world semantics, therefore, connects the notion of *necessity* and *possibility* to the universal and existential quantification ($\forall, \exists$) under first-order logic.

[4] This confounding factor affects the examples in Table 9 of Han et al. (2024). See Bertolazzi et al. (2024) for a detailed discussion on the impact of knowledge bias in natural language reasoning.

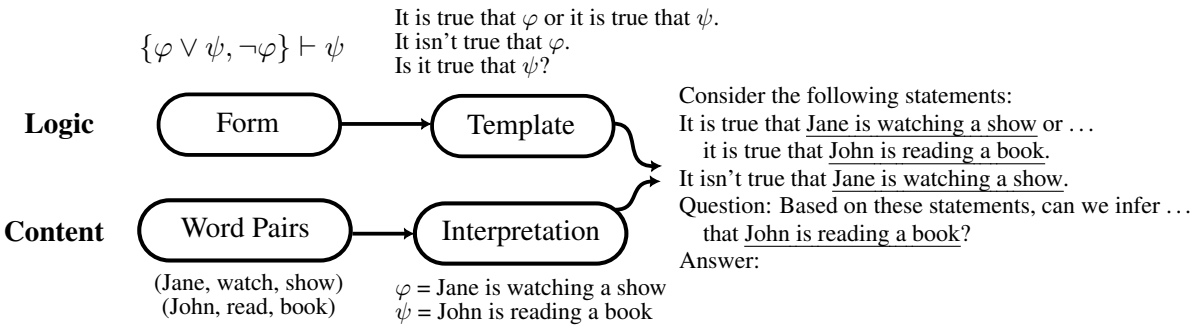

Figure 2: The data synthesis pipeline: for each variable in logic forms (§3.1), we assign meanings to them to obtain the natural language question-answering pairs (§3.2).

premises. To mitigate the ambiguity in natural language, we design heuristic rules to translate logic forms into less ambiguous English, which are detailed in Appendix B.1. For the exact wordings we used, see Table A1 in Appendices. If a valid deduction exists ($\vdash$) for the logical form, the ground truth answer is `Yes`, otherwise `No`. The answer is solely determined by the logical form and is independent of the interpretation.

3.3 Involved Logical Forms

Translated logical forms can have varying degrees of naturalness. For example, the *necessitation rule* $\{\varphi\} \vdash \Box\varphi$, which translates to "$\varphi$ is true; therefore, it is certain that φ is true," appears to be unnatural due to redundancy. [5] Based on the relationship between $\vee$ and $\rightarrow$ in Eq. (2), we use hypothetical and disjunctive syllogisms with four basic variants:

$$\{\varphi \vee \psi, \neg\varphi\} \vdash \psi, \qquad (\vee^{\mathrm{L}})$$
$$\{\neg\varphi \rightarrow \psi, \neg\varphi\} \vdash \psi, \qquad (\rightarrow^{\mathrm{L}}\text{; modus ponens})$$
$$\{\varphi \vee \psi, \neg\psi\} \vdash \varphi, \qquad (\vee^{\mathrm{R}})$$
$$\{\neg\varphi \rightarrow \psi, \neg\psi\} \vdash \varphi. \qquad (\rightarrow^{\mathrm{R}}\text{; modus tollens})$$

Despite the semantic similarity, these logical forms translate to different natural-language questions. For example, taking the interpretations of φ := *Jane is watching a show* and ψ := *John is reading a book*, $\vee^{\mathrm{L}}$ translates to

> Consider the following statements:
> *Jane is watching a show or John is reading a book.*
> *Jane isn't watching a show.*
> Question: Based on these statements, can we infer that *John is reading a book*?

[5] Nevertheless, we report the experiment results on necessitation rule in Appendix C.1.

With the same interpretation, $\rightarrow^{\mathrm{L}}$'s translation of the first statement is *If Jane isn't watching a show, then John is reading a book.*

According to the commutativity of disjunction operator, we group $\vee^{\mathrm{L}}$ and $\vee^{\mathrm{R}}$ together as *disjunctive syllogism*, alongside two hypothetical syllogism groups, modus ponens ($\rightarrow^{\mathrm{L}}$) and modus tollens ($\rightarrow^{\mathrm{R}}$). All the logical forms shown above are valid sequents with ground-truth answer `Yes`. To balance the dataset, we introduce some logic fallacies that generate questions with ground-truth label `No`. By flipping the second premises and the conclusions, we obtain the following fallacies:

$$\{\varphi \vee \psi, \psi\} \nvdash \neg\varphi, \qquad (\vee^{\mathrm{L}}_{\not\vdash})$$
$$\{\neg\varphi \rightarrow \psi, \psi\} \nvdash \neg\varphi, \qquad (\rightarrow^{\mathrm{L}}_{\not\vdash})$$
$$\{\varphi \vee \psi, \varphi\} \nvdash \neg\psi, \qquad (\vee^{\mathrm{R}}_{\not\vdash})$$
$$\{\neg\varphi \rightarrow \psi, \varphi\} \nvdash \neg\psi, \qquad (\rightarrow^{\mathrm{R}}_{\not\vdash})$$

where $\vee^{\mathrm{L}}_{\not\vdash}$ and $\vee^{\mathrm{R}}_{\not\vdash}$ are grouped as *affirming the disjunction*, $\rightarrow^{\mathrm{L}}_{\not\vdash}$ and $\rightarrow^{\mathrm{R}}_{\not\vdash}$ corresponds to *affirming the consequent* and *denying the antecedent*, respectively. In our dataset, we require the formulas φ and ψ to the form of $\mathfrak{M}p$ and $\mathfrak{M}q$, where p and q are propositional variables, each assigned with an interpretation. Both variables are constrained under the same modality $\mathfrak{M}$, which can be necessity ($\Box$), possibility ($\Diamond$) or no modality ($\varnothing$). Pairing with four rules and theorem–fallacy variations, we have a total of $3 \times 4 \times 2 = 24$ forms.

3.4 Involved Logic Interpretations

For logic interpretations, we generate a set of verb phrases by prompting the CodeLlama 2 model (Rozière et al., 2024), and select 204 of them manually. and combine them with top-200 popular baby names in the US into subject-verb-object pairs,[6]

[6] https://www.ssa.gov/oact/babynames/names.zip

such as (*Ray*, *make*, *a pizza*). We randomly generate 1000 interpretations with two pairs each. The same set of interpretations is applied to variables p, q in each logic sequent's natural langauge template. In total, there are $24 \times 1000 = 24000$ question, with samples shown in Table A1.

4 Experiment

4.1 Metrics and Investigated Models

Hu and Levy (2023) have suggested that the standard approach of greedily decoding yes-no strings (Dentella et al., 2023) may underestimate the competence of a language model; therefore, we adopt a probability-based metric to evaluate the model performance. In our evaluation protocol, the predicted likelihood of the tokens `Yes` and `No`, conditioned on the prompt s—denoted as $p(\texttt{Yes} \mid s)$ and $p(\texttt{No} \mid s)$, respectively—serve as the soft labels for yes-no answers. The soft accuracy $\hat{p}$ on the single example with ground-truth answer $y \in \{\texttt{Yes}, \texttt{No}\}$ is defined as the relative probability of y:

$$\hat{p} = \frac{p(\texttt{No} \mid s)\mathbb{1}[y = \texttt{No}] + p(\texttt{Yes} \mid s)\mathbb{1}[y = \texttt{Yes}]}{p(\texttt{No} \mid s) + p(\texttt{Yes} \mid s)},$$

where $\mathbb{1}[\cdot]$ is the indicator function that returns 1 if the condition is true and 0 otherwise. This relative probability can also be viewed as the confidence score of the model on the ground-truth answer. The soft accuracy Acc_{soft} of a model on the entire dataset $\mathcal{D}$ is defined as the average soft accuracy over all examples,

$$Acc_{soft} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\hat{p}_i.$$

We use a zero-shot setting to investigate the general performance of the models' logical inference capabilities—while adding detailed instructions or few-shot demonstrations may increase the absolute performance, they are at the cost of introducing possibly undesired confounding factors or behaviors, such as simply copy-pasting the answers in the examples.

We evaluate on the following models with open-sourced weights: mistral-7b-v0.2 and -8x7b (Jiang et al., 2023, 2024); llama-2-7b, -13b and -70b (Touvron et al., 2023); 3.1 version of llama-3-8b and -70b (AI@Meta, 2024); yi-34b (01.AI, 2024); phi-2 and phi-3-mini (Microsoft, 2023, 2024).[7]

[7]Our evaluation protocol technically requires the conditional probabilities of specified answers given a prompt, which are not supported by most commercial models; however, we report the greedy-decoding accuracy of these models on a sample subset for reference.

4.2 Results: Performance w.r.t. Logical Forms

We evaluate the aforementioned models with the probability-based protocol (Table 1). Generally, models that rank higher in the leaderboard also achieve higher soft accuracy on our dataset. The break-down accuracies on modalities and argument forms reveal that:

1. (Modality) All models consistently perform better on the possibility ($\Diamond$) than necessity ($\Box$) or plain propositional logic.
2. (Argument Forms) The pattern is more diverse, yet most of the models struggle the most on modus tollens ($\rightarrow^{\text{R}}_{\not\vdash}$) within logic sequents (i.e., questions with ground-truth answers `Yes`), and affirming the consequent ($\rightarrow^{\text{L}}_{\not\dashv}$) within fallacies. This result resonates with Wason (1968), who concludes that modus tollens is a particularly hard reasoning pattern for human participants with card selection tasks.

4.2.1 Analysis on Logic Sequents

To systematically analyze the effect on model performance of each factor of interest, as well as cross-validating the observations above, we fit a linear mixed-effects model (Raudenbush, 2002) to the soft accuracy data on valid logic sequents (i.e. with ground truth of `Yes`) across different LLMs and logical forms,

$$\begin{aligned} Acc_{soft} \sim\ & Modality + ArgForm + Perplexity \\ & + (1 + Perplexity \mid LLM), \end{aligned} \tag{4}$$

with the linear fixed effects of (i.) modality, (ii.) argument form, and (iii.) input perplexity. Individual probability, coupled with a constant term, is modeled as a random effect to account for potential model-specific biases. Here, *Perplexity* denotes the perplexity of the input text $(x_1 x_2 \ldots x_N)$, which is defined as the exponential of the token-wise average negative log-likelihood of the text given a specific language model:

$$Perplexity = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(x_i \mid x_{<i})\right)$$

The mixed-effects model yields a marginal R^2 of 0.342 and a conditional R^2 of 0.543, suggesting

Model	Overall (Rank)		Leaderboard (Rank)		Modality ∅	Modality □	Modality ◇	Argument Form $\vee^{L,R}_{\vdash}$	Argument Form $\rightarrow^{L}_{\vdash}$	Argument Form $\rightarrow^{R}_{\vdash}$	Argument Form $\vee^{L,R}_{\nvdash}$	Argument Form $\rightarrow^{L}_{\nvdash}$	Argument Form $\rightarrow^{R}_{\nvdash}$
mistral-7b	0.645	(4)	0.145	(7)	0.464	0.496	**0.974**	0.877	0.663	0.280	0.434	0.653	**0.939**
mistral-8x7b	0.724	(1)	0.193	(5)	0.698	0.601	**0.874**	**0.963**	0.873	0.023	0.757	0.648	0.813
llama-2-7b	0.335	(10)	0.094	(10)	0.262	0.207	**0.538**	0.444	0.147	0.315	0.208	0.451	**0.468**
llama-2-13b	0.513	(9)	0.110	(9)	0.488	0.362	**0.688**	0.418	0.581	0.393	**0.631**	0.436	0.591
llama-2-70b	0.611	(5)	0.127	(8)	0.616	0.471	**0.746**	0.446	**0.845**	0.518	0.775	0.389	0.694
llama-3-8b	0.565	(6)	0.239	(3)	0.598	0.460	**0.639**	0.526	0.470	0.332	0.664	0.625	**0.716**
llama-3-70b	0.714	(2)	0.362	(1)	0.745	0.554	**0.843**	0.606	0.773	0.515	**0.882**	0.661	0.788
yi-34b	0.518	(8)	0.226	(4)	0.457	0.413	**0.683**	0.346	0.498	0.205	0.685	0.638	**0.737**
phi-2	0.532	(7)	0.155	(6)	0.469	0.456	**0.673**	0.670	**0.757**	0.522	0.365	0.402	0.510
phi-3-mini	0.690	(3)	0.272	(2)	0.657	0.536	**0.877**	0.839	**0.974**	0.475	0.664	0.462	0.604
OpenAI-o1	0.926	N/A	N/A	N/A	**1.000**	0.773	**1.000**	0.895	**1.000**	0.775	0.919	**1.000**	**1.000**
Gemini-1.5-Pro	0.859	N/A	N/A	N/A	0.831	0.748	0.997	**1.000**	**1.000**	0.919	0.661	0.991	0.638
human	0.595	N/A	N/A	N/A	0.589	0.566	**0.640**	0.691	**0.901**	0.628	0.594	0.225	0.411

Table 1: Overall and break-down accuracies of different models, as well as their HuggingFace OpenLLM Leaderboard performance and relative ranking (Fourrier et al., 2024). Each argument form category denotes the union of the fine-grained categories specified in the superscripts and subscripts—for example, $\vee^{L,R}_{\vdash}$ denotes the entire disjunctive syllogism group. **Boldfaced** values indicate the row-wise maximum for each factor. Note that due to technical limitations of commercial LLMs, results from OpenAI-o1 (OpenAI, 2024) and Gemini-1.5-pro (Team et al., 2024) are greedy-decoding based evaluation on 2,000 random samples that serve as references, and are therefore not directly comparable to other probability-based evaluations. Human results are detailed in §5.

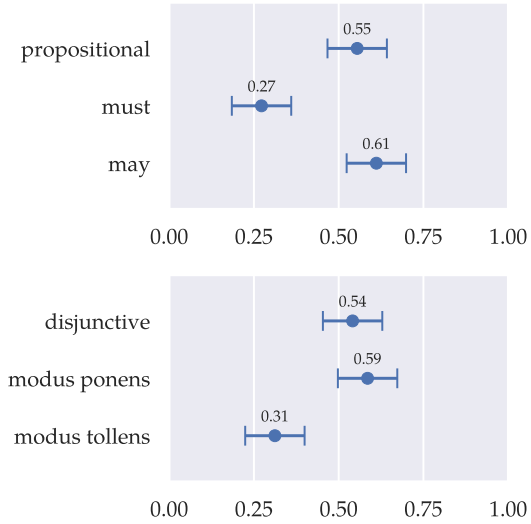

Figure 3: Estimated marginal means of logical form factors in the mixed-effects model of Eq. (4), along with their 95% confidence intervals.

Hypothesis	p**-value**
propositional < may	< 0.001
must < propositional	< 0.001
must < may	< 0.001
disjunctive < modus ponens	< 0.001
modus tollens < modus ponens	< 0.001
modus tollens < disjunctive	< 0.001

Table 2: Hypothesis testing results on the effect of logical form factors on soft accuracy (Figure 3).

a reasonable predictive power. The likelihood ratio test on the full regression model vs. the null regression model without each of the fixed effects yields a significant result ($p < 0.001$), suggesting the importance of all these factors in determining the model performance.

Fixed effects. In line with Gonen et al. (2023) and McCoy et al. (2024), we find a negative correlation between perplexity and soft accuracy ($p < 0.001$); however, the correlation is weak ($\rho = -0.09$), which suggests the necessity of the complementary factors below in predicting LLM performance.

For different modalities and argument forms, we estimate their marginal means on soft accuracy (Figure 3) and perform pairwise hypothesis testing on the estimated coefficients (Table 2). The results generally align with the general observations on the full dataset. The only exception is that modus ponens ($\rightarrow^{L}$), instead of disjunctive syllogisms ($\vee$), appears to be the easiest argument form (i.e., the one with the highest soft accuracy) among all.

Random effects. We analyze the per-LLM random effects on the soft accuracy (Figure 4). All the model-specific mixed effects of perplexity are negative, suggesting the negative correlation between perplexity and soft accuracy is consistent across models (Figure 4a). While the intercept random effects are not perfectly aligned with the model performance—since the perplexity random

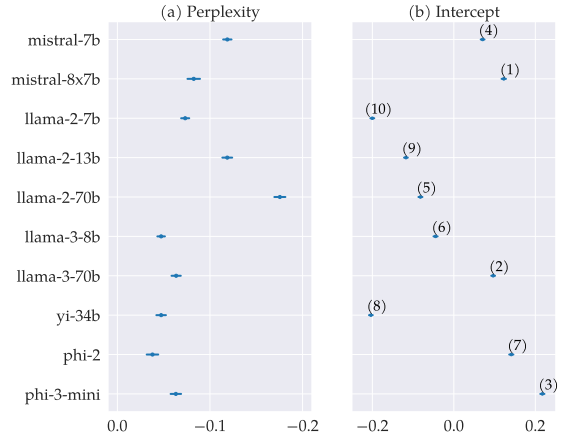

Figure 4: Illustration of per-model random effects on soft accuracy in the mixed-effects model of Eq. (4) with 99.9% confidence intervals. (a) Mixed effects (i.e., the sum of fixed and random effects) of perplexity. (b) Intercept random effects (i.e., constant term per model on soft accuracy), with the model performance rank (Table 1) annotated in parentheses.

effects may introduce confounding factors—higher-ranked models generally tend to have higher intercept random effects (Figure 4b), which cross-validates the general performance ranking.

4.2.2 Extended Analysis on the Negative Perplexity–Performance Correlation

We further investigate the negative correlation between perplexity and model performance through a controlled experiment: we create a mirror dataset of the same size, keeping all the logical formulas while interpreting them with nonsensical words. For example, the formula $\Diamond(\varphi \vee \psi)$ may be interpreted as *it's possible that Neva is balaring a montery or Lucille is sweeling prandates*, where the underlined words and phrases are nonsensical. Intuitively, the perplexity of the problems in this mirror dataset should be much higher than that of the primary dataset problems (§3) under any reasonably trained language model.

We analyze the correlation between perplexity and model performance (Figure 5). As desired, the perplexity of problems with nonsensical words are indeed much higher than that of the primary dataset ($\approx$ 30 vs. $\approx$ 10). The significant portion of horizontal and inclined lines in the figures again suggests that perplexity is not a reliable predictor of model performance. Meanwhile, the overall parallelism of the lines echos our results that logical forms are important factors for such prediction.

4.2.3 The Affirmation Bias over Modalities

One key argument of Dentella et al. (2023) is that large language models exhibit a bias towards affirming the claim, i.e., answering `Yes` more frequently than `No`. We investigate this phenomenon by fitting a mixed-effects model

$$\frac{P(\texttt{Yes} \mid s)}{P(\texttt{Yes} \mid s) + P(\texttt{No} \mid s)} \sim \mathit{Modality} + \mathit{ArgForm} + \mathit{Perplexity} + (1 + \mathit{Perplexity} \mid \mathit{LLM}), \quad (5)$$

which has the same structure as Eq. (4), except the dependent variable being the relative probability of answering `Yes` conditioned on input text s.

We present the estimated marginal means of the factors in the mixed-effects model (Figure 6). While our results confirm the affirmation bias on propositional logic, such bias is slightly less pronounced on the possibility modality ($\Diamond$, around 0.03), and the models even show a bias towards rejecting claims under the necessity modality ($\Box$).

5 Human Experiments

LLMs are trained on text produced by humans and are able to generate plausible text; therefore, there have been interests in using LLMs as human models (Eisape et al., 2024; Misra and Kim, 2024, *inter alia*). Following this line of work, we conduct a human behavioral experiment to ground the LLM reasoning behavior, as advocated in Ivanova (2025). Using samples from our primary dataset, we collected 710 responses from adults fluent in English through Prolific.[8] More experiment details can be found in Appendix A.2.

The average human accuracy on each group is shown in the last row of Table 1.[9] Aligned with our LLM results (§4), on modalities, the overall human results also show an accuracy order of ($\Diamond \succ \varnothing \succ \Box$), and on argument forms, modus ponens ($\rightarrow^{\mathrm{L}}$) is the most accurately answered pattern.

To further investigate the interactions of logic factors, we fit a generalized linear mixed-effects model (Bates et al., 2015) to verify the effect of modality and argument forms on human logic reasoning accuracy (Eq. (6) and Figure 7).

$$\mathrm{logit}(\mathit{Acc}) \sim \mathit{Modality} + \mathit{ArgForm} + \mathit{Rt} + (1 + \mathit{Rt} \mid \mathit{ParticipantID}), \quad (6)$$

[8] https://prolific.com

[9] Human responses are binary classes, so correct and incorrect responses are coded as 1 and 0, respectively.

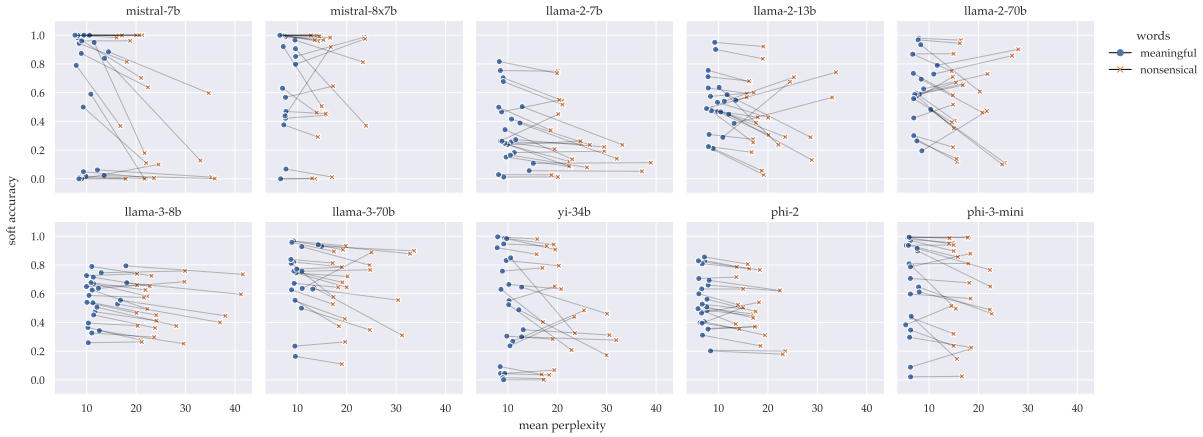

Figure 5: Correlation between mean perplexity and mean confidence score on each logic sequent. Each point represents an average over a group of 1000 prompts that share the same underlying logic sequent. Two connected dots share the same logic formula.

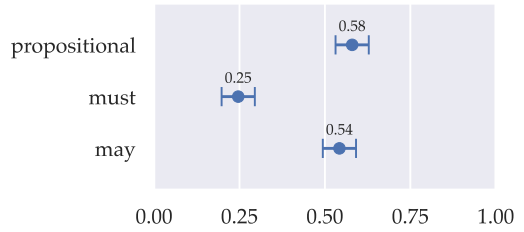

Figure 6: Estimated marginal means of the factors in the mixed-effects model of Eq. (5) with 95% confidence intervals. Higher coefficients indicate a higher tendency to affirm the claim.

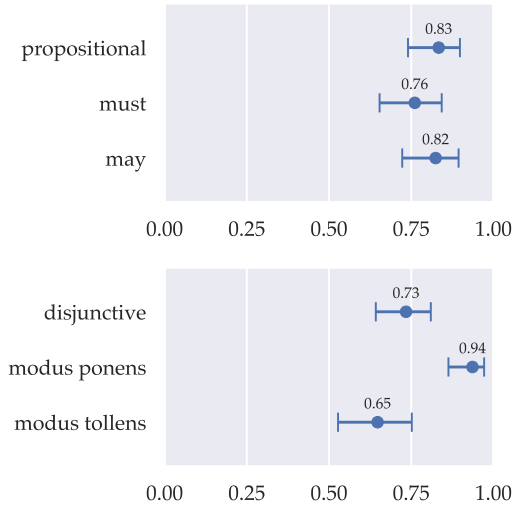

Figure 7: Estimated marginal means of logical form factors in the generalized mixed-effects model of Eq. (6), along with their 95% confidence intervals.

where Acc is the binary accuracy of human responses, and Rt is the response time. The generalized mixed-effects model yields a marginal R^2 of 0.121 yet a 0.419 conditional R^2, indicating a diverse response pattern across participants. The likelihood ratio test on the full model against the null model shows that only the effect of argument form is significant ($\chi^2(2) = 25.6$, $p < 0.001$). However, in accordance with the overall performance, we find modus ponens ($\rightarrow^{\mathrm{L}}$) has a significantly higher effect than the other two valid argument forms. This confirms that logical forms can also have a significant impact on human reasoning accuracy, which is consistent with the LLM results, although the effect sizes are not the same.

6 Conclusion and Discussion

We present an analysis of hypothetical and disjunctive syllogisms on propositional and modal logic and systematically analyze the LLM performance on the dataset. Our analysis provides novel insights on explaining and predicting LLM performance: in addition to the perplexity or probability of the input text, the underlying logic forms play an important role in determining the performance of LLMs. In addition, we compare the behaviors of LLMs and humans using the same data through human behavioral experiments. We discuss the implications of our results as follows.

Probability in language models. Probability and perplexity are often used as intrinsic evaluation metrics for language models. While Gonen et al. (2023) and McCoy et al. (2024) show that probability and perplexity correlate well with LLM performance, literature in program synthesis with LLMs shows little correlation between probability and execution-based evaluation results (Li et al., 2022; Shi et al., 2022). This work does not necessarily contradict either line but rather provides complementary factors for analyzing LLM performance.

We argue that probability may have become an

overloaded term in analyzing LLMs. Low probability may be due to one or more of the following non-exhaustive reasons: (1) out-of-context content, (2) ungrammatical language, or (3) grammatical but semantically awkward content (cf. the mirror dataset in §4.2.2), (4) reasonable but rare content. We hypothesize that the probability of language models may not be essentially able to capture all these nuanced differences, and call for encoding and decoding algorithms—such as Meister et al. (2023)—that can better decompose the probability into finer-grained and explainable components.

Comparing humans and LLMs. What is our goal for building LLMs? To achieve better performance on practical tasks or to build a more human-like model? Our results, together with Eisape et al. (2024), suggest that these two goals may not be perfectly aligned by revealing a mixture of similarity and discrepancy between LLMs and humans—for example, while LLMs exhibit higher benchmark performance than humans on our dataset and show the same argument form preferences with humans (Figures 3 and 7), they also show systematic biases that we do not find significant in human reasoning (e.g., disfavoring the necessity modality, §4.2.3). While there has been positive evidence of using LLMs as human models in psycholinguistic studies (Misra and Kim, 2024, *inter alia*), our results suggest executing such approaches cautiously.

On the relation between modality and performance. Our results show that there is a significant difference in performance between necessity and possibility modalities, with the former much lower than the latter (Table 1). Part of the reason for this is that LLMs have a significant tendency to say "No" to the necessity modality (Figure 6).

On the one hand, our results extend the conclusion of Dentella et al. (2023) that LLMs generally respond positively—LLM behaviors may be significantly affected by finer-grained factors, including but not necessarily limited to the modality involved in the input. On the other hand, while LLMs systematically tend to answer "No" to questions in necessity modality, we do not find related evidence in human experiments, which leads us to hypothesize that such rejection bias comes from either the model architecture or the training strategies, such as the reinforcement learning with human feedback (RLHF; Ouyang et al., 2022) protocol. We leave this as an open question for future research.

Modal logic and theory of mind. Modality, in principle, encodes mental states and beliefs. The reasoning of beliefs also resonates with the theory of mind (Premack and Woodruff, 1978; Baron-Cohen et al., 1985, *inter alia*) and machine theory of mind (Rabinowitz et al., 2018; Ma et al., 2023, *inter alia*). Following the effort by Sileo and Lernould (2023) that uses epistemic modal logic to model the machine theory of mind, our work assesses the behaviors of LLMs on alethic modal logic, distantly revealing the future potential of LLMs in achieving the theory of mind.

Limitations

This work comes with two major limitations:

1. While we have verified that our data has a low perplexity (9.82 ± 2.47 under mistral-7b; much lower than that of the data by Wan et al. (2024), 25.44), and, therefore, are similar enough to natural language utterances, the synthetic language cannot fully substitute natural language in daily life. Our dataset and analysis are not comprehensive enough to cover many nuanced examples that may appear in real communication, especially when context-dependent understanding is crucial to conveying communication goals.

2. Despite more than 7,000 languages worldwide, as a first step, our material only covers English. This narrow focus is due to the languages the authors are proficient in and the coverage of the language models. We acknowledge the importance of extending the scope of this work to a more comprehensive set of languages and leave the extension as an immediate follow-up step.

In addition, the sample size of human experiments is somewhat limited. We leave more comprehensive human behavioral data collection and analysis to future work.

Ethics Statement

While this work involves human logical reasoning experiments, we have ensured that (1) the data are generated procedurally following templates listed in the paper and (2) there is no harmful content in the atomic logical interpretations, reviewed by all the authors. In addition, we have ensured that all participants are paid a fair wage through the Prolific platform. Instructions and consent forms delivered to the participants can be found in the Appendix A.2. The institutional ethics review board has approved the data collection process.

This work contributes to the understanding of LLMs. We do not foresee risk beyond the minimal risk posed by LLM evaluation work. We acknowledge that using LLMs in real-world scenarios could significantly impact human behaviors, raising the need for model transparency, safety, security, and interpretability. We will open-source the synthetic logical reasoning dataset upon publication.

7 Acknowledgements

We thank Yudong Li for his help in setting up the Gemini and OpenAI API for the experiments. This work was supported in part by a Google PhD Fellowship and a Canada CIFAR AI Chair award to FS, as well as NSERC RGPIN-2024-04395.

References

01.AI. 2024. Yi: Open Foundation Models by 01.AI.

AI@Meta. 2024. The Llama 3 Herd of Models.

Roberta Ballarin. 2023. Modern origins of modal logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2023 edition. Metaphysics Research Lab, Stanford University.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1).

Catarina G. Belem, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of Linguistic Uncertainty by Language Models and Humans.

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A Systematic Analysis of Large Language Models as Soft Reasoners: The Case of Syllogistic Inferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13882–13905. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as soft reasoners over language. In *IJCAI*.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *NAACL*.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of ACL: EMNLP*.

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

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: Natural Language Reasoning with First-Order Logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Wesley H. Holliday, Matthew Mandelkern, and Cedegao E. Zhang. 2024. Conditional and Modal Reasoning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3800–3821, Miami, Florida, USA. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *ICML*, pages 9118–9147. PMLR.

Anna A. Ivanova. 2025. How to evaluate the cognitive abilities of LLMs. 9(2):230–233.

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.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard University Press.

Henry A Kautz, Bart Selman, et al. 1992. Planning as satisfiability. In *ECAI*, volume 92, pages 359–363. Citeseer.

Saul A. Kripke. 1959. A Completeness Theorem in Modal Logic. *The Journal of Symbolic Logic*, 24(1):1–14.

Saul A. Kripke. 1963. Semantical Analysis of Modal Logic I Normal Modal Propositional Calculi. *Mathematical Logic Quarterly*, 9(5-6):67–96.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3622–3628, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. *Findings of Empirical Methods in Natural Language Processing*.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Microsoft. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Microsoft. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models: A case study on cross-dative generalization. *arXiv preprint arXiv:2408.05086*.

Santiago Ontanon, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. 2022. LogicInference: A new Datasaet for Teaching Logical Inference to seq2seq Models. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*.

OpenAI. 2024. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2024. Exploring Reasoning Biases in Large Language Models Through Syllogism: Insights from the NeuBAROCO Dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16063–16077. Association for Computational Linguistics.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4218–4227. PMLR.

Marco Ragni, Hannah Dames, Daniel Brand, and Nicolas Riesterer. 2019. When Does a Reasoner Respond: Nothing Follows?: 41st Annual Meeting of the Cognitive Science Society. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 2640–2645.

Stephen W Raudenbush. 2002. Hierarchical linear models: Applications and data analysis methods. *Advanced Quantitative Techniques in the Social Sciences Series/SAGE*.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code Llama: Open Foundation Models for Code.

Abulhair Saparov and He He. 2022. 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. *Advances in Neural Information Processing Systems*, 36:3083–3105.

S Seals and Valerie Shalin. 2024. Evaluating the Deductive Competence of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8614–8630. Association for Computational Linguistics.

Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I Wang. 2022. Natural language to code translation with execution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Stuart M. Shieber. 1993. The problem of logical form equivalence. *Computational Linguistics*, 19(1):179–190.

Damien Sileo and Antoine Lernould. 2023. MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *EMNLP*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Louis Martin, and Kevin Stone. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. 2024. LogicAsker: Evaluating and Improving the Logical Reasoning Ability of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2155, Miami, Florida, USA. Association for Computational Linguistics.

Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281.

Yimei Xiang. 2019. Two types of higher-order readings of wh-questions. *Proceedings of the 22nd Amsterdam Colloquium*.

Shi Zong and Jimmy Lin. 2024. Categorical syllogisms revisited: A review of the logical reasoning abilities of llms for analyzing categorical syllogism. *arXiv preprint arXiv:2406.18762*.

A Additional Experiment Details

A.1 LLM Experiment Details

All LLMs used are obtained from Hugging Face checkpoints, exact repository and commits are shown in Table 3. Specifically, we use the instruction-tuned versions of the models due to the zero-shot nature of our experiments.

Time and compute power requirements vary, the largest llama-3-70b model takes around 2 hours on NVIDIA A6000 GPU to obtain all results in §4.

All models besides `phi-2` and `phi-3-mini` is quantized to the bfloat16 floating-point format (bf16), while the Phi family is using the original 32-bit format (f32). We found that `phi-2`'s performance degraded too much when quantized in an earlier experiment, so we resort to f32 on small models from the Phi family. Due to compute resource limits, models of size 70B have to be quantized, so we choose to align all the larger models to bf16.

A.2 Human Experiment Details

Participant instructions. We use keys *F* and *J*, which are roughly symmetric on a standard English keyboard, to collect participant responses. Half of the participants see the following instruction:

In this study, you will be presented with two statements followed by a question. Your task is to answer either Yes or No to the question, based on the information provided in the statements. Please respond quickly and accurately by pressing "F" for Yes, and "J" for No.

Model	Repository	Commit
mistral-7b	mistralai/Mistral-7B-Instruct-v0.2	250544c9a802b0396550d0fd24bc80ff98bb1f5f
mistral-8x7b	mistralai/Mixtral-8x7B-Instruct-v0.1	1e637f2d7cb0a9d6fb1922f305cb784995190a83
llama-2-7b	meta-llama/Llama-2-7b-chat-hf	f5db02db724555f92da89c216ac04704f23d4590
llama-2-13b	meta-llama/Llama-2-13b-chat-hf	c2f3ec81aac798ae26dcc57799a994dfbf521496
llama-2-70b	meta-llama/Llama-2-70b-chat-hf	8b17e6f4e86be78cf54afd49ddb517d4e274c13f
llama-3-8b	meta-llama/Meta-Llama-3.1-8B-Instruct	5206a32e0bd3067aef1ce90f5528ade7d866253f
llama-3-70b	meta-llama/Meta-Llama-3.1-70B-Instruct	33101ce6ccc08fa6249c10a543ebfcac65173393
yi-34b	01-ai/Yi-34B-Chat	75fa0ab8d2a50841f433f373433c170b571638ba
phi-2	microsoft/phi-2	ef382358ec9e382308935a992d908de099b64c23
phi-3-mini	microsoft/Phi-3-mini-4k-instruct	5a516f86087853f9d560c95eb9209c1d4ed9ff69

Table 3: Hugging Face checkpoints used in the experiments.

To mitigate the possible bias introduced by the dominant hand, we have the other half of the participants see instruction with reversed keys:

In this study, you will be presented with two statements followed by a question. Your task is to answer either Yes or No to the question, based on the information provided in the statements. Please respond quickly and accurately by pressing "F" for No, and "J" for Yes.

Participant wage. We offer participants an hourly wage of 1.5 times Prolific's minimum wage. The duration is determined by the median completion time among all participants.

B Extra Details of the Dataset

B.1 Considerations in Translating Logical Form to Natural Language

During the interpretation process, another key point is to assign independent interpretations to variables. Deciding the dependency also involves common sense knowledge. For example, consider the premises $\neg p \to q$ and q. If we interpret $p :=$ *"Jane is inside the house"* and $q :=$ *"Jane is out"* to proposition variables p and q, the two variables are possibly not independent. According to common sense, "*Jane is not inside the house*" ($\neg p$) correlates with or is even equivalent to "*Jane is out*" (q). Logically, $\{\neg p \to q, q\} \nvdash \neg p$; however, with the extra premise $\neg p \leftrightarrow q$ given by common sense, people may conclude that $\neg p$.[10]

Besides, natural language is ambiguous—one sentence in natural language can come from multiple logical forms under the same interpretation. We use present tense and progressive aspect to encourage a reading of imaginary ongoing events, corresponding to the alethic modality. Such events are less likely to induce LLM's or human's individual bias, as they are unrelated to factual knowledge or moral judgements. Also, we always use two full verb phrases, ruling out sentences like "*Jane is eating apples or oranges,*" so the two events are less likely to be mutually exclusive. In this way, we can reduce the ambiguity of the questions in our dataset.

B.2 Data Samples

All logic forms and corresponding natural language sentences can be found in Table A1.

The exact prompt format is as follows:

```
Consider the following statements:\n
Jane is watching a show or John is reading a book.\n
Jane isn't watching a show.\n
Question: Based on these statements, can we infer that
John is reading a book?\n
Answer:<eof>
```

C Additional Experiments

C.1 Extra Experiment: Introduction Rule of Modality

We report the results on the necessitation rule and its variants here, as these rules are obscure and verbose to be articulated in natural language:

$$\{\varphi\} \vdash \Box\varphi, \qquad \text{(necessitation rule)}$$
$$\{\varphi\} \vdash \Diamond\varphi,$$
$$\{\varphi\} \vdash \varphi.$$

Its natural language form is as follows:

Jane is watching a show.
- ($\Box$) Can we infer that it's certain that Jane is watching a show?
- ($\Diamond$) Can we infer that it's possible that Jane is watching a show?
- ($\varnothing$) Can we infer that Jane is watching a show?

All three variants are paired with 1000 logic interpretations. As they are all rules of inference,

[10] This confounding factor affects the examples in Appendix C.1.12 of Holliday et al. (2024).

Modality	Argument Form	Logical Form	Natural Language
∅	$\vee^{\mathrm{L}}$	$\{p \vee q, \neg p\} \vdash q$	Jane is watching a show or John is reading a book. Jane isn't watching a show. Can we infer that John is reading a book?
∅	$\vee^{\mathrm{R}}$	$\{p \vee q, \neg q\} \vdash p$	Jane is watching a show or John is reading a book. John isn't reading a book. Can we infer that Jane is watching a show?
∅	$\rightarrow^{\mathrm{L}}$	$\{\neg p \rightarrow q, \neg p\} \vdash q$	If Jane isn't watching a show, then John is reading a book. Jane isn't watching a show. Can we infer that John is reading a book?
∅	$\rightarrow^{\mathrm{R}}$	$\{\neg p \rightarrow q, \neg q\} \vdash p$	If Jane isn't watching a show, then John is reading a book. John isn't reading a book. Can we infer that Jane is watching a show?
□	$\vee^{\mathrm{L}}$	$\{\Box p \vee \Box q, \neg \Box p\} \vdash \Box q$	It's certain that Jane is watching a show or it's certain that John is reading a book. It's uncertain whether Jane is watching a show. Can we infer that it's certain that John is reading a book?
□	$\vee^{\mathrm{R}}$	$\{\Box p \vee \Box q, \neg \Box q\} \vdash \Box p$	It's certain that Jane is watching a show or it's certain that John is reading a book. It's uncertain whether John is reading a book. Can we infer that it's certain that Jane is watching a show?
□	$\rightarrow^{\mathrm{L}}$	$\{\neg \Box p \rightarrow \Box q, \neg \Box p\} \vdash \Box q$	If it's uncertain whether Jane is watching a show, then it's certain that John is reading a book. It's uncertain whether Jane is watching a show. Can we infer that it's certain that John is reading a book?
□	$\rightarrow^{\mathrm{R}}$	$\{\neg \Box p \rightarrow \Box q, \neg \Box q\} \vdash \Box p$	If it's uncertain whether Jane is watching a show, then it's certain that John is reading a book. It's uncertain whether John is reading a book. Can we infer that it's certain that Jane is watching a show?
◇	$\vee^{\mathrm{L}}$	$\{\Diamond p \vee \Diamond q, \neg \Diamond p\} \vdash \Diamond q$	It's possible that Jane is watching a show or it's possible that John is reading a book. It's impossible that Jane is watching a show. Can we infer that it's possible that John is reading a book?
◇	$\vee^{\mathrm{R}}$	$\{\Diamond p \vee \Diamond q, \neg \Diamond q\} \vdash \Diamond p$	It's possible that Jane is watching a show or it's possible that John is reading a book. It's impossible that John is reading a book. Can we infer that it's possible that Jane is watching a show?
◇	$\rightarrow^{\mathrm{L}}$	$\{\neg \Diamond p \rightarrow \Diamond q, \neg \Diamond p\} \vdash \Diamond q$	If it's impossible that Jane is watching a show, then it's possible that John is reading a book. It's impossible that Jane is watching a show. Can we infer that it's possible that John is reading a book?
◇	$\rightarrow^{\mathrm{R}}$	$\{\neg \Diamond p \rightarrow \Diamond q, \neg \Diamond q\} \vdash \Diamond p$	If it's impossible that Jane is watching a show, then it's possible that John is reading a book. It's impossible that John is reading a book. Can we infer that it's possible that Jane is watching a show?
∅	$\vee^{\mathrm{L}}$	$\{p \vee q, q\} \nvdash \neg p$	Jane is watching a show or John is reading a book. John is reading a book. Can we infer that Jane isn't watching a show?
∅	$\vee^{\mathrm{R}}$	$\{p \vee q, p\} \nvdash \neg q$	Jane is watching a show or John is reading a book. Jane is watching a show. Can we infer that John isn't reading a book?
∅	$\rightarrow^{\mathrm{L}}$	$\{\neg p \rightarrow q, q\} \nvdash \neg p$	If Jane isn't watching a show, then John is reading a book. John is reading a book. Can we infer that Jane isn't watching a show?
∅	$\rightarrow^{\mathrm{R}}$	$\{\neg p \rightarrow q, p\} \nvdash \neg q$	If Jane isn't watching a show, then John is reading a book. Jane is watching a show. Can we infer that John isn't reading a book?
□	$\vee^{\mathrm{L}}$	$\{\Box p \vee \Box q, \Box q\} \nvdash \neg \Box p$	It's certain that Jane is watching a show or it's certain that John is reading a book. It's certain that John is reading a book. Can we infer that it's uncertain whether Jane is watching a show?
□	$\vee^{\mathrm{R}}$	$\{\Box p \vee \Box q, \Box p\} \nvdash \neg \Box q$	It's certain that Jane is watching a show or it's certain that John is reading a book. It's certain that Jane is watching a show. Can we infer that it's uncertain whether John is reading a book?
□	$\rightarrow^{\mathrm{L}}$	$\{\neg \Box p \rightarrow \Box q, \Box q\} \nvdash \neg \Box p$	If it's uncertain whether Jane is watching a show, then it's certain that John is reading a book. It's certain that John is reading a book. Can we infer that it's uncertain whether Jane is watching a show?
□	$\rightarrow^{\mathrm{R}}$	$\{\neg \Box p \rightarrow \Box q, \Box p\} \nvdash \neg \Box q$	If it's uncertain whether Jane is watching a show, then it's certain that John is reading a book. It's certain that Jane is watching a show. Can we infer that it's uncertain whether John is reading a book?
◇	$\vee^{\mathrm{L}}$	$\{\Diamond p \vee \Diamond q, \Diamond q\} \nvdash \neg \Diamond p$	It's possible that Jane is watching a show or it's possible that John is reading a book. It's possible that John is reading a book. Can we infer that it's impossible that Jane is watching a show?
◇	$\vee^{\mathrm{R}}$	$\{\Diamond p \vee \Diamond q, \Diamond p\} \nvdash \neg \Diamond q$	It's possible that Jane is watching a show or it's possible that John is reading a book. It's possible that Jane is watching a show. Can we infer that it's impossible that John is reading a book?
◇	$\rightarrow^{\mathrm{L}}$	$\{\neg \Diamond p \rightarrow \Diamond q, \Diamond q\} \nvdash \neg \Diamond p$	If it's impossible that Jane is watching a show, then it's possible that John is reading a book. It's possible that John is reading a book. Can we infer that it's impossible that Jane is watching a show?
◇	$\rightarrow^{\mathrm{R}}$	$\{\neg \Diamond p \rightarrow \Diamond q, \Diamond p\} \nvdash \neg \Diamond q$	If it's impossible that Jane is watching a show, then it's possible that John is reading a book. It's possible that Jane is watching a show. Can we infer that it's impossible that John is reading a book?

Table A1: Samples of all logical forms and corresponding natural language sentences.

	∅	□	◇
mistral-7b	0.998	0.885	0.999
mistral-8x7b	0.957	0.540	0.987
llama-2-7b	0.768	0.013	0.920
llama-2-13b	0.368	0.004	0.829
llama-2-70b	0.511	0.051	0.834
llama-3-8b	0.398	0.225	0.783
llama-3-70b	0.674	0.384	0.794
yi-34b	0.960	0.382	0.999
phi-2	0.814	0.226	0.892
phi-3-mini	0.992	0.925	0.994

Table A2: Overall accuracy of the necessitation rule and its modality variants on each model.

Modality	Argument Form	Logical Form
∅	base	$\varphi \vee \psi, \neg\varphi \vdash \psi$
□	base	$\Box\varphi \vee \Box\psi, \neg\Box\varphi \vdash \Box\psi$
□	theorem	$\Box(\varphi \vee \psi), \Box\neg\varphi \vdash \Box\psi$
□	spurious	$\Box(\varphi \vee \psi), \neg\Box\varphi \nvdash \Box\psi$
◇	base	$\Diamond\varphi \vee \Diamond\psi, \neg\Diamond\varphi \vdash \Diamond\psi$
◇	theorem	$\Diamond(\varphi \vee \psi), \Diamond\neg\varphi \vdash \Diamond\psi$
◇	spurious	$\Diamond(\varphi \vee \psi), \neg\Diamond\varphi \vdash \Diamond\psi$

Table A3: Logical forms and their ground truth to study the distribution of modalities. Only the spurious form of the necessity modality (marked by underline) has a ground truth of false.

the ground truth answer is always `Yes`. Overall accuracy is shown in Table A2, where across all LLMs, the necessitation rule has the lowest accuracy. This echoes the necessity modality's tendency to be rejected discussed in §4.2.3.

We further fit a linear mixed-effects model similar to Eq. (4), except that the argument form effect is now constant across all data points. The mixed-effects model yields a marginal R^2 of 0.391 and a conditional R^2 of 0.745. Estimated marginal means shows that the accuracy on ∅ is 0.171 less than ◇, but 0.371 higher than □, with both differences significant at $p < 0.0001$. This further suggests that modality serves as an important factor on logic reasoning performance.

C.2 Extra Experiment: Distribution of Modalities

Besides the necessitation rule, *distribution axiom* is the other fundamental axiom in normal modal logic. It can be transformed into the rule shown in Eq. (A1), and plugging in the definition of $\vee$ in Eq. (2) gives the rule shown in Eq. (A2). Notice that Eq. (A2) closely resembles rule $\vee^{\mathrm{L}}$'s variant with necessity, as shown in Eq. (A3), except the different scope of the necessity operator and the position of the negation operator. Moving the negation operator out of the necessity operator will result in a fallacy (Eq. A4).

$$\{\Box(\varphi \to \psi), \Box\varphi\} \vdash \Box\psi, \tag{A1}$$
$$\{\Box(\varphi \vee \psi), \Box\neg\varphi\} \vdash \Box\psi, \tag{A2}$$
$$\{\Box\varphi \vee \Box\psi, \neg\Box\varphi\} \vdash \Box\psi, \tag{A3}$$
$$\{\Box(\varphi \vee \psi), \neg\Box\varphi\} \nvdash \Box\psi. \tag{A4}$$

We say (A2) to (A4) are of argument form `theorem`, `base` and `spurious`, respectively. See Table A3 for the logical forms and their ground truth we used to study the distribution of modalities. The natural language form is as follows:

	It's certain that if Freddy is not going shopping, then Coy is making dinner.
`(theorem)`	It's certain that Freddy is not going shopping.
`(spurious)`	It's uncertain whether Freddy is going shopping.
	Can we infer that it's certain that Coy is making dinner?

This group of rules and fallacies comes from the fact that the necessity modality □ is not distributive to disjunction, i.e. $\Box(\varphi \vee \psi) \nvdash \Box\varphi \vee \Box\psi$ (Xiang, 2019, Ex. 5). In contrast, the possibility modality ◇ is distributive to disjunction. This particular case could have served as a material to test the LLM's knowledge of the asymmetry between the two modalities, yet in §4.2.3 we showed that there is a bias towards rejection on the necessity modality. As the false case of the disjunction is on the necessity modality, this bias confounds the experiment.

We fit a linear mixed-effects model similar to Eq. (4) to the data,

$$\begin{aligned} \mathit{Acc}_{\mathit{soft}} \sim\ & \mathit{Modality} \times \mathit{ArgForm} + \mathit{Perplexity} \\ & + (1 + \mathit{Perplexity} \mid \mathit{LLM}), \end{aligned}$$

with an interaction term between the modality and argument form. On the `theorem` form compared to the `base` form, the necessity modality □ has a 0.173 higher estimated marginal means with $p < 0.0001$ significance, yet the possibility modality ◇ has a 0.071 lower estimated marginal means. On the `spurious` form compared to the `base` form, the □ has a 0.312 higher means, and the ◇ has no significant difference. On both forms, $\Diamond \succ \Box$ in terms of accuracy still holds at a slight margin of 0.110 and 0.047 respectively.

	Soft	Soft (Thresholded)	Hard
mistral-7b	0.644	0.648	0.648
mistral-8x7b	0.725	0.731	0.730
llama-2-7b	0.335	0.273	0.259
llama-2-13b	0.513	0.553	0.511
llama-2-70b	0.611	0.719	0.696
llama-3-8b	0.566	0.684	0.665
llama-3-70b	0.714	0.850	0.836
yi-34b	0.517	0.544	0.528
phi-2	0.532	0.544	0.544
phi-3-mini	0.690	0.712	0.712

Table A4: Comparison between greedy decoding-based hard accuracy and probability-based soft accuracy.

To verify whether on □ the performance increase on `spurious` form is due to the rejection bias, we fit a linear mixed-effects model with the relative probability of answering `Yes` as dependent variable. Results show that on `spurious` form compared to the `base` form, the effect of □'s tendency to answer `Yes` is only 0.060 lower, indicating the rejection bias of the `base` form is still present. Therefore, we hypothesize that the LLM's performance on recognizing the fallacy of necessity distribution over disjunction is hindered by the rejection bias on the necessity modality.

C.3 Extra Experiment: Greedy Decoding

We provide additional results based on greedy decoding methods. The hard accuracy, a metric complementing the soft accuracy we mentioned in §4.1, is obtained by the following procedure:

1. Prefill the input.
2. Predict the next token with the model, and greedily select the token t with the highest probability.
3. If the t token is `Yes` or `No`, take the token as model's answer and terminate.
4. Append the t token to the input. Return to step 2 to generate autoregressively, until an answer is determined or the maximum number of output tokens N is reached.
5. The hard accuracy is 1 iff the answer is correct. If the answer is incorrect or the maximum number of output tokens N is reached without an answer, the hard accuracy is 0.

We report the hard accuracy with maximum token number of $N = 10$. [11] As hard accuracy is a binary metric, we threshold the soft accuracy at 0.5 ($\mathbb{I}[Acc_{soft} \geqslant 0.5]$) for a fair comparison. The results of accuracies averaged over all argument forms and modalities on all models are shown in Table A4. We observe that the thresholded soft accuracy is no less than the hard accuracy on models. Besides accuracy, we also observe that although nearly all models always predict an answer token (either `Yes` or `No`) as the first output token, this is not the case for `phi-3-mini`. It on average greedily decodes an answer token at around the 4th output token, and only on 76.3% of the cases it decodes an answer token within the first 10 output tokens. The evidence above supports the argument from Hu and Levy (2023) that decoding-based methods may underestimate the language model's true performance, and justifies our choice of using the soft accuracy in §4.

[11] Notice that due to the non-deterministic nature of GPU parallelism, the soft accuracy results reported in this section differ from the one in §4. We verify that on all models, the mean squared error between the soft accuracy results reported here and the ones in §4 is less than 0.01.