Normative Reasoning in Large Language Models: A Comparative Benchmark from Logical and Modal Perspectives

Kentaro Ozeki[1,2], Risako Ando[1], Takanobu Morishita[1], Hirohiko Abe[1], Koji Mineshima[1], Mitsuhiro Okada[1]
[1]Keio University, Tokyo, Japan
[2]University of Tokyo, Tokyo, Japan
kentaro.ozeki@gmail.com {risakochaan,morishita,hirohiko-abe}@keio.jp
{minesima,okada}@abelard.flet.keio.ac.jp

Abstract

Normative reasoning is a type of reasoning that involves normative or deontic modality, such as obligation and permission. While large language models (LLMs) have demonstrated remarkable performance across various reasoning tasks, their ability to handle normative reasoning remains underexplored. In this paper, we systematically evaluate LLMs' reasoning capabilities in the normative domain from both logical and modal perspectives. Specifically, to assess how well LLMs reason with normative modals, we make a comparison between their reasoning with normative modals and their reasoning with epistemic modals, which share a common formal structure. To this end, we introduce a new dataset covering a wide range of formal patterns of reasoning in both normative and epistemic domains, while also incorporating non-formal cognitive factors that influence human reasoning. Our results indicate that, although LLMs generally adhere to valid reasoning patterns, they exhibit notable inconsistencies in specific types of normative reasoning and display cognitive biases similar to those observed in psychological studies of human reasoning. These findings highlight challenges in achieving logical consistency in LLMs' normative reasoning and provide insights for enhancing their reliability. All data and code are released publicly at https://github.com/kmineshima/NeuBAROCO.

1 Introduction

Recent research and development of large language models (LLMs) has placed increasing emphasis on their reasoning capabilities, particularly in tasks such as mathematical problem-solving and coding (Mahowald et al., 2024). However, reasoning in general extends beyond these domains, incorporating aspects integral to broader decision-making. One such area is *normative reasoning*, which involves normative or deontic modality, such as obligation, permission, and prohibition (negative obligation).

P: It is **obligatory** to answer your question.	P: It is not **permitted** to answer your question.
C: It is **permitted** to answer your question.	C: It is not **obligatory** to answer your question.
pattern: $\mathsf{O}A \Rightarrow \mathsf{P}A$	**pattern**: $\neg\mathsf{P}A \Rightarrow \neg\mathsf{O}A$
expected: valid	**expected**: valid

Figure 1: The two reasoning patterns are logically related (contrapositive) but LLMs often struggle to make consistent predictions. $\mathsf{O}A$ means "A is obligatory" and $\mathsf{P}A$ means "A is permitted." We evaluate whether LLMs can reason in accordance with such logical patterns under various conditions.

Normative reasoning can be viewed as a specialized facet of social reasoning, which has recently received substantial attention in LLM research (Mondorf and Plank, 2024; Mahowald et al., 2024; Almeida et al., 2024). Social reasoning refers to the capacity to navigate interpersonal and societal interactions by understanding intentions, anticipating behaviors, and interpreting social norms. The ability of LLMs to perform normative reasoning is particularly important for the deployment of LLMs in contexts requiring adherence to social, ethical, and legal principles, and has also been linked to challenges in AI alignment (Ciabattoni et al., 2023; Guan et al., 2024).

Normative reasoning involves formal and contentual aspects. Prior research on normative reasoning in LLMs has primarily focused on contentual aspects, such as cultural and social factors embedded in the content that influence model behavior (Sheng et al., 2021; Navigli et al., 2023). By contrast, although there is growing interest in the logical or formal reasoning abilities of LLMs (Clark et al., 2020; Bertolazzi et al., 2024; Cheng et al., 2025), the logical aspect of normative reasoning in LLMs has received less attention. A language model capable of reliable normative reasoning should consistently recognize and ap-

ply valid reasoning patterns (Figure 1). This leads to the question of the extent to which LLMs can demonstrate logical consistency in normative reasoning.

In this paper, we systematically evaluate the capabilities of LLMs in normative reasoning, in terms of logical validity and invalidity. To achieve this, we design reasoning tasks that assess LLMs' consistency across formal reasoning patterns as well as the influence of non-formal factors, extending prior studies to the normative domain in two main directions. First, to highlight the distinctive characteristics of LLMs in normative reasoning, we compare their performance with another type of modal reasoning, *epistemic reasoning*. Epistemic reasoning involves logical inference with epistemic modals, which concern a reasoner's knowledge and beliefs, a domain in which LLMs have been shown to exhibit gaps in basic reasoning (Holliday et al., 2024). Second, we compare the LLM reasoning with existing findings from cognitive psychology on human biases in normative reasoning, thereby extending results on LLMs' reasoning biases in general (Ando et al., 2023; Lampinen et al., 2024; Ozeki et al., 2024). To support these evaluations, we introduce a new dataset designed for a comparative benchmark of normative and epistemic reasoning in LLMs, while also incorporating non-formal cognitive factors known to influence human reasoning.

Our key findings are summarized as follows:

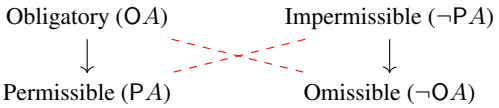
- Even the best-performing models often show inconsistencies in basic normative reasoning, such as the inference from obligation to permission (illustrated in Figure 1, left).
- The models also manifest human-like biases, such as content effects, in both normative and epistemic reasoning.
- While it is commonly assumed in the cognitive science literature that normative reasoning is easier than epistemic reasoning for humans, due to domain-specific aspects of cognition (Section 3.1), our findings show that language models do not necessarily follow this pattern: their relative performance in the two domains varies across tasks.
- Reasoning involving negation is particularly challenging for the models, as evidenced by performance on the Syllogistic Task. The presence of negation has a greater impact on reasoning difficulty than the distinction between entailment (valid) and non-entailment (invalid), a contrast previously discussed in the literature (Eisape et al., 2024; Ozeki et al., 2024).

2 Patterns of Normative Reasoning

By normative reasoning, we refer to reasoning with deontic modals such as obligation, permission, and prohibition. In this study, we focus on two basic types of normative reasoning: (1) **Deontic Logic Reasoning**, which involves single-premise logical inferences and is used to evaluate basic understanding of modal concepts, including challenges specific to modality of obligation, such as Ross's paradox and the free choice inference; and (2) **Syllogistic Reasoning**, which involves multi-premise logical inferences and incorporates normative rules and generalizations, including patterns involving universal quantification and conditional statements.

2.1 Deontic Logic Reasoning

Deontic logic is a formal theory of normative reasoning (von Wright, 1951; Gabbay et al., 2013). In deontic logic, obligations such as "It is obligatory that A" and "One must do A" are symbolized as $\mathsf{O}A$, while permissions such as "It is permissible that A" and "One may do A" are symbolized as $\mathsf{P}A$. Obligation and permission are analyzed as deontic necessity and possibility, where $\neg\mathsf{O}A$ ("It is not obligatory to do A") is equivalent to $\mathsf{P}\neg A$ ("It is permitted not to do A") and $\neg\mathsf{P}A$ ("It is not permitted to do A") is equivalent to $\mathsf{O}\neg A$ ("It is obligatory not to do A"). Obligation and permission, when combined with negation, follow a traditional inference scheme known as "The Deontic Square" (McNamara and Van De Putte, 2022).



For example, an inference from $\mathsf{O}A$ ("It is obligatory that A") to $\mathsf{P}A$ ("It is permissible that A") is *logically valid* in the sense that if the premise is true, the conclusion is also true. This is indicated by the arrow from $\mathsf{O}A$ to $\mathsf{P}A$ in the square. Similarly, an inference from $\neg\mathsf{P}A$ ("It is not permissible that A") to $\neg\mathsf{O}A$ ("It is not obligatory that A") is also logically valid. The dotted lines in the square indicate that these pairs are a contradiction.

Label	Form	Premise Example	Hypothesis Example
NotMu-MiNot	$\neg \mathsf{O}A \Rightarrow \mathsf{P}\neg A$	It is not mandatory to take a shower every day.	It is acceptable not to take a shower every day.
NotMi-MuNot	$\neg \mathsf{P}A \Rightarrow \mathsf{O}\neg A$	You are not permitted to litter.	It is mandatory not to litter.
MiNot-NotMu	$\mathsf{P}\neg A \Rightarrow \neg \mathsf{O}A$	It is permissible not to attend the party.	There is no obligation to attend the party.
Mu-Mi	$\mathsf{O}A \Rightarrow \mathsf{P}A$	You must take care of your health.	You can choose to take care of your health.
NotMi-NotMu	$\neg \mathsf{P}A \Rightarrow \neg \mathsf{O}A$	It is not acceptable to lie in court.	It is not the case that you must lie in court.
NotMu-NotMi	$\neg \mathsf{O}A \Rightarrow \neg \mathsf{P}A$	You are not required to use the internet.	You are not allowed to use the internet.
MiNot-MuNot	$\mathsf{P}\neg A \Rightarrow \mathsf{O}\neg A$	You are allowed not to drive a car.	You must not drive a car.
Mi-Mu	$\mathsf{P}A \Rightarrow \mathsf{O}A$	It is permissible to help others.	You are required to help others.
FC-Or-Elim	$\mathsf{P}(A \vee B) \Rightarrow \mathsf{P}A$	You may travel to Japan or France.	You may travel to Japan.
FC-Or-Intro	$\mathsf{P}A \Rightarrow \mathsf{P}(A \vee B)$	You may learn to sing.	You may learn to sing or dance.
Ross-Or-Intro	$\mathsf{O}A \Rightarrow \mathsf{O}(A \vee B)$	You must tell the truth.	You must tell the truth or lie.

Table 1: 11 patterns of basic deontic logic reasoning. "$\neg$" denotes negation (not), "$\vee$" denotes disjunction (or), and "$\phi \Rightarrow \psi$" represents an inference from the premise ϕ to the hypothesis ψ. Mi and Mu in the label refer to permission and obligation, respectively, while NotMi and NotMu denote negations of permission and obligation. MiNot and MuNot denote permission and obligation of negations. Those in blue are *valid* patterns, while those in red are *invalid* patterns.

Label: Cat-MP P1: All B must C P2: A is B C: A must C	**Label**: Cat-MT P1: All B must C P2: A is not required to C C: A is not B	**Label**: Cat-AC P1: All B must C P2: A must C C: A is B	**Label**: Cat-DA P1: All B must C P2: A is not B C: A is not required to C
Label: Hyp-MP P1: If A is B, A must C P2: A is B C: A must C	**Label**: Hyp-MT P1: If A is B, A must C P2: A is not required to C C: A is not B	**Label**: Hyp-AC P1: If A is B, A must C P2: A must C C: A is B	**Label**: Hyp-DA P1: If A is B, A must C P2: A is not B C: A is not required to C

Table 2: 8 patterns of normative (deontic) syllogisms. P1: Major Premise, P2: Minor Premise, C: Conclusion; Cat: Categorical, Hyp: Hypothetical; MP: Modus Ponens, MT: Modus Tollens, AC: Affirming the Consequent, DA: Denying the Antecedent. B and C represent predicates, and A represents a term. Those in blue are *valid* patterns, while those in red are *invalid* patterns.

Table 1 presents all formal patterns of deontic logic reasoning examined in the present work.

All foregoing inference patterns are valid within Standard Deontic Logic (SDL). However, it is well known that SDL fails to adequately capture certain intuitively valid or invalid patterns of deontic reasoning. For instance, consider the inference from "You may eat an apple or a banana" to "You may eat an apple" ($\mathsf{P}(A \vee B) \Rightarrow \mathsf{P}A$). While this inference is intuitively plausible, it is not valid in SDL, as it relies on disjunction elimination, a rule invalid under the standard interpretation of disjunction: from $A \vee B$ one cannot infer A, and SDL inherits this property. This phenomenon is known as the Free Choice paradox (Kamp, 1973), labeled FC-Or-Elim in Table 1.

Conversely, suppose the inference from "You must post the letter" to "You must post the letter or burn it" ($\mathsf{O}A \Rightarrow \mathsf{O}(A \vee B)$). Although intuitively questionable, this inference is valid in SDL, due to the classical logical principle that A entails $A \vee B$. This discrepancy is known as Ross's paradox (Ross, 1941), labeled Ross-Or-Intro in Table 1.

These paradoxes reveal a fundamental tension between logical validity in SDL and human judgments in normative reasoning. In particular, they raise the question of how disjunctions are interpreted in normative contexts: do models follow logical behavior, in which "$A \vee B$ does not entail A" and "A entails $A \vee B$"? Or do they exhibit modality-specific reasoning, where "$\mathsf{P}(A \vee B)$ entails $\mathsf{P}A$" (FC-Or-Elim), and "$\mathsf{O}A$ does not entail $\mathsf{O}(A \vee B)$" (Ross-Or-Intro), thereby aligning with intuitive interpretations of normative language?

One of the central goals of our evaluation of deontic reasoning is to test which of these patterns LLMs follow: whether they behave in accordance with standard logical disjunction, or exhibit reasoning patterns that are characteristic of deontic modality and human intuition.

2.2 Syllogistic Reasoning

As a form of inference involving multiple premises, we focus on *syllogisms*, building on prior relevant studies (Ando et al., 2023; Lampinen et al., 2024; Eisape et al., 2024; Ozeki et al., 2024). Syllogism is the type of logical reasoning drawing a conclusion (**C**) from two premises (**P1**, **P2**). We consider

Label	Form	Premise	Hypothesis
NotMu-MiNot	$\neg\Box A \Rightarrow \Diamond\neg A$	It is not certain that the economy will recover quickly.	The economy might not recover quickly.
NotMi-MuNot	$\neg\Diamond A \Rightarrow \Box\neg A$	It is not possible that a person can read a 500-page book in one hour.	It is certain that a person cannot read a 500-page book in one hour.
MiNot-NotMu	$\Diamond\neg A \Rightarrow \neg\Box A$	It is possible that the theory will not be proven wrong.	It is not known that the theory will be proven wrong.
Mu-Mi	$\Box A \Rightarrow \Diamond A$	It is established that the Earth's climate is changing.	There's a chance that the Earth's climate is changing.
NotMi-NotMu	$\neg\Diamond A \Rightarrow \neg\Box A$	It is not possible that a person can perfectly recall every event in their life.	It is not certain that a person can perfectly recall every event in their life.
NotMu-NotMi	$\neg\Box A \Rightarrow \neg\Diamond A$	It is not certain that a person's personality is fixed at birth.	It is not possible that a person's personality is fixed at birth.
MiNot-MuNot	$\Diamond\neg A \Rightarrow \Box\neg A$	It is possible that the painting is not authentic.	It is known that the painting is not authentic.
Mi-Mu	$\Diamond A \Rightarrow \Box A$	There is a possibility that the universe is teeming with life.	Life must have been teeming in the universe.

Table 3: 8 patterns of *epistemic* logic reasoning matched to those of normative logic reasoning. Mi and Mu in the label refer to epistemic possibility and epistemic necessity, respectively, while NotMi and NotMu denote negations of epistemic possibility and epistemic necessity. MiNot and MuNot denote epistemic possibility and epistemic necessity of negations.

cases in which one of the premises (**P1**) expresses a normative rule or generalization. When **P1** is a universally quantified statement, the inference is called *categorical syllogism* (Cat); when **P1** takes the form of an "if-then" conditional sentence, the inference is called a *hypothetical syllogism* (Hyp).

To examine whether the presence or absence of negation affects the difficulty of inference, we further classify syllogisms into four distinct patterns: Modus Ponens (MP), Modus Tollens (MT), Affirming the Consequent (AC), and Denying the Antecedent (DA). The following is an instance of a categorical syllogism with the Modus Tollens pattern (Cat-MT), a logically valid form of inference.

P1: All first-year students are required to submit assignments.
P2: Mia is not required to submit assignments.
C: Mia is not a first-year student.

As an example of an invalid syllogism, the following is a hypothetical syllogism with Denying the Antecedent pattern (Hyp-DA):

P1: If Mia is a first-year student, then she must submit assignments.
P2: Mia is not a first-year student.
C: Mia is not required to submit assignments.

In these patterns, negation appears in both **P2** and **C**, while the inference is logically valid in Cat-MT but invalid in Hyp-DA. This contrast provides a useful test case for evaluating how the presence of negation and the distinction between valid and invalid reasoning affect the difficulty of inference.

There are a total of eight inference patterns, all of which are presented in Table 2. Each sentence (involving normative modality) in normative syllogisms is expressed in various expressions, as shown in Table 8 in Appendix A.

3 Non-Formal Factors

To examine how normative reasoning is influenced by factors beyond formal patterns, we consider the effects of modality (domain specificity) and inference content (content effects).

3.1 Domain Specificity

A key question in the study of normative reasoning is whether it exhibits distinctive characteristics that set it apart from other types of reasoning. An influential view in cognitive science holds that reasoning is domain-specific, meaning that reasoning in different domains is governed by distinct cognitive mechanisms. This is known as *domain specificity* of reasoning (Cosmides, 1989; Cosmides and Tooby, 1992; Fiddick, 2004). Seals and Shalin (2024) report that LLMs perform better in reasoning with social content than with non-social content, mirroring human reasoning, though the effect is less prominent in LLMs than in humans. However, it remains unclear whether this domain-specific reasoning ability extends to normative reasoning as in the case of humans (Cheng and Holyoak, 1985).

We investigate this question by comparing normative reasoning with *epistemic reasoning*. Similar to deontic logic for normative reasoning, the formal framework developed to model epistemic reasoning is known as *epistemic logic* (Hintikka, 1962; Rendsvig et al., 2024). Epistemic logic in-

Content Type	Description	Example Sentences
Congruent	Premise and conclusion align with common sense.	Deontic: *It is not obligatory to eat breakfast* Epistemic: *It is not certain that AI will replace all jobs*
Incongruent	Premise or conclusion contradicts common sense.	Deontic: *It is not obligatory to care for your children* Epistemic: *It is not certain that fire needs oxygen*
Nonsense	Sentences use nonsensical or made-up words.	Deontic: *It is not obligatory to flibbertigibbet* Epistemic: *It is not certain that the flooglehorp grimples the zizzle.*

Table 4: Three types of content used to analyze content effects in deontic and epistemic reasoning.

cludes two types of modal expressions reflecting necessity and possibility in terms of knowledge and beliefs. **Epistemic necessity** indicates that something is certain given the evidence, and includes expressions such as *It is known that...*, *It is certain that...*, or the modal *must* (e.g., *She must be an expert*). **Epistemic possibility** indicates that something may be true given the reasoner's knowledge and beliefs, and includes expressions such as *It is possible that...* or the modal *might* (e.g., *She might be an expert*).

Some inference patterns that are valid in deontic logic are also valid in epistemic logic, and the same holds for invalid patterns. For example, the following is an epistemic version of an invalid hypothetical syllogism with Denying the Antecedent pattern (`Hyp-DA`), whose deontic counterpart we saw in Section 2.2.

P1: If a student misses orientation, then they must be new to the program.
P2: Sam did not miss orientation.
C: Sam must not be new to the program.

All patterns of epistemic logic reasoning and epistemic syllogisms examined in this study are listed in Table 3 and in Table 7 in Appendix A, respectively. Variations in the expressions used to represent epistemic modality are shown in Table 9 in Appendix A.

In the literature, Holliday et al. (2024) provide a systematic evaluation of LLMs' reasoning with epistemic modality. The present study extends this line of work by investigating how the modality of reasoning—deontic versus epistemic—affects LLM behavior, through a comparative analysis using the two tasks described above.

3.2 Content Effects

The *content effect* is the human tendency to accept inferences whose conclusions align with one's beliefs and to reject those whose conclusions do not, regardless of their logical validity (Evans et al., 1993). Recent studies have shown that LLMs exhibit similar behavior: they tend to make reasoning errors when conclusions contradict common-sense beliefs (Ando et al., 2023; Lampinen et al., 2024; Ozeki et al., 2024).

To analyze content effects in deontic reasoning, we categorize each inference into one of three content types based on the nature of its premise and conclusion. These categories distinguish whether the statements are consistent with common-sense knowledge, contradict it, or are nonsensical. The three types are summarized in Table 4.

4 Experiments

In this section, we describe the data and task formats, the models used in the experiments, and the evaluation settings.

4.1 Data and Task Formats

We evaluate two types of inference problems, **Deontic Logic** and **Syllogistic**, as described in Section 2. Each type is divided into normative reasoning and epistemic reasoning problems using various modal vocabularies as explained in Section 3.1. All problems are written in English.

For the Deontic Logic task, we manually created 11 templates for normative inference (see Table 1) and 8 templates for epistemic inferences (see Table 3). For the Syllogistic task, we created 8 templates for normative and epistemic inferences (see Table 2). We then instantiated these templates with 20 concrete words for each of the three content types explained in Section 3.2, using Gemini 1.5 Pro, which was not used in the evaluation. Finally, we manually refined the resulting instances, when necessary, to ensure consistency and quality.

The Deontic Logic task includes 640 problems for normative reasoning (360 valid, 280 invalid) and 480 problems for epistemic reasoning (300

valid, 180 invalid). The Syllogistic task includes 480 problems (240 valid and 240 invalid) for normative and epistemic reasoning, respectively.

4.2 Models

For our analysis, we evaluated 5 recently released language models, using their instruction-tuned versions when available. GPT-4o and GPT-4o-mini (OpenAI, 2024) are state-of-the-art closed-weight models, accessed via the OpenAI API. Llama-3.1-8B-In (8B parameters) and Llama-3.3-70B-In (70B parameters) are instruction-tuned versions of open-weight Llama 3 models (Dubey et al., 2024). Phi-4 is a 14B-parameter open-weight model that achieves strong reasoning performance relative to its size (Abdin et al., 2024).

4.3 Evaluation

The models are evaluated based on the accuracy of their predictions against the expected answers. We conduct experiments using three types of prompts. In the Zero-Shot setting, the prompt contains only the instruction for the task and the problem. In the Few-Shot setting, exemplars for in-context learning are included (Brown et al., 2020). Here, we provide a single exemplar with expected answers for each reasoning pattern, where the patterns differ depending on whether the problem belongs to the Normative or Epistemic domain. In the Chain-of-Thought (CoT) setting, the model is prompted to generate a sequence of reasoning steps that lead to the final conclusion (Wei et al., 2022). Specifically, we adopt the Zero-Shot CoT approach (Kojima et al., 2022).

The evaluation is performed in a single run with the temperature set to 0.0 to make the model responses deterministic. When the expected output is a single word, we limit the maximum output tokens to 10. In the CoT setting, the maximum output token is extended to 1024. Other hyperparameters are kept at their default values.

5 Results and Analysis

In this section, we present the results of the normative reasoning tasks. Table 5 presents the overall performance of the models on the Normative problems in the two tasks, in comparison to those on the Epistemic problems. GPT-4o in the Few-Shot setting achieves the highest accuracy among the models in the Normative domain on all tasks. Among the smaller models, Phi-4 shows competitive performance to the larger models.

5.1 Comparison by Prompt Type

The performance of the models is generally higher in the Few-Shot setting compared to the Zero-Shot setting. However, Llama models show a decrease in performance with Few-Shot prompts on the Syllogistic task. Chain-of-Thought prompting contributes to a small improvement or has negative impact in the Deontic Logic task, while it has a larger impact on the Syllogistic task.

5.2 Consistency across Reasoning Patterns

First, to investigate the consistency of the models across reasoning patterns of different forms, we compare the performance of the models on the different normative reasoning tasks. The complete results and error examples in the Deontic Logic and Syllogistic tasks are presented in Appendix C.

Deontic Logic Task. Figure 2 illustrates the performance of the best-performing model (GPT-4o) on the most challenging patterns in the Deontic Logic task: `Mu-Mi` (the inference from obligation to permission) and the paradox-related inferences (`FC-Or-Intro` and `Ross-Or-Intro`). Overall, the models perform well when the patterns align with Standard Deontic Logic. However, in the `Mu-Mi` pattern, most models exhibit lower performance, with Llama-3.3-70B-In being the only model to perform well. For the controversial patterns, namely `FC-Or-Elim`, `FC-Or-Intro`, and `Ross-Or-Intro`, the models demonstrate mixed performance. They perform well on the `FC-Or-Elim`, where valid is the expected answer that aligns with common-sense reasoning. In contrast, LLMs tend to accept the validity of `FC-Or-Intro` and `Ross-Or-Intro`, which are expected to be invalid. Regarding the relative difficulty of valid and invalid patterns, no clear trend is observed.

Syllogistic Task. Figure 3 presents the performance of the best-performing model (GPT-4o) in the Syllogistic task. Both in the categorical (`Cat`) and hypothetical (`Hyp`) cases, the models perform well on the `MP` and `AC` patterns, but show lower performance on the `MT` and in particular `DA` patterns. The latter two involve negation in the premises and hypotheses, whereas the former two do not. Previous studies have shown that LLMs struggle with reasoning involving negation (Truong et al., 2023; García-Ferrero et al., 2023). Our results in the Syllogistic task corroborate this observation.

Model	Deontic Logic						Syllogistic					
	Normative			Epistemic			Normative			Epistemic		
	Zero	Few	CoT	Zero	Few	CoT	Zero	Few	CoT	Zero	Few	CoT
gpt-4o-mini	81.41	87.19	80.47	84.79	85.21	82.08	69.58	82.08	76.91	63.33	72.29	76.56
gpt-4o	84.22	97.03	85.31	93.75	92.29	91.67	78.33	92.29	85.90	62.92	83.96	76.88
llama-3.1-8B-In	70.31	73.28	73.28	76.04	86.83	67.29	56.25	46.67	58.33	57.08	52.50	54.79
llama-3.3-70B-In	78.75	94.06	80.78	91.67	91.25	95.63	56.25	46.67	58.33	57.08	52.50	54.79
phi-4	82.50	80.94	78.59	92.08	84.79	88.33	75.62	67.29	80.83	75.21	76.25	66.25

Table 5: Overall accuracy (%) for Normative and Epistemic problems across the tasks. Prompting strategies: **Zero** = Zero-Shot, **Few** = Few-Shot, **CoT** = Chain-of-Thought. Shading follows a gradient from red (0%) to blue (100%), with white representing the midpoint (50%).

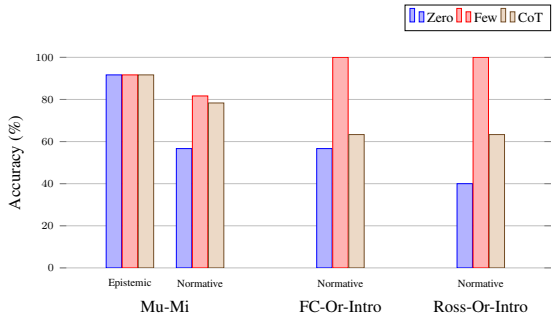


Figure 2: Accuracy (%) of the best-performing model (GPT-4o) for Mu-Mi, FC-Or-Intro, Ross-Or-Intro patterns in the Deontic Logic task. The FC-Or-Intro and Ross-Or-Intro patterns are Normative problems only.

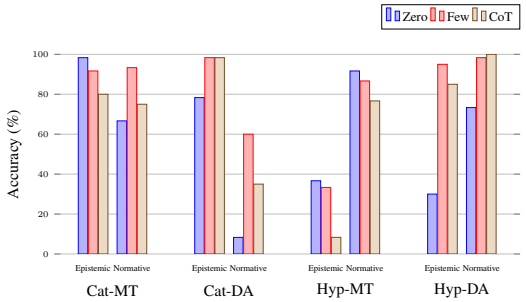


Figure 3: Accuracy (%) of the best-performing model (GPT-4o) for Cat-MT, Cat-DA, Hyp-MT, Hyp-DA patterns in the Syllogistic task.

5.3 Non-Formal Factors

Domain Specificity. To assess whether models perform better on normative or epistemic reasoning, we compare their performance across both domains. The overall results for Normative and Epistemic domains are presented in Table 5. In the Syllogistic task, models perform comparably or better on normative problems than on epistemic ones under the Zero-Shot setting, which aligns with human tendency studied in cognitive science. In contrast, in the Deontic Logic task, normative problems are more challenging than epistemic problems under the Zero-Shot setting. These results suggest that the relationship between reasoning domain and model performance may vary across tasks and reasoning patterns.

(a) Deontic Logic

Model	Incong.	Cong.	Nonsense
gpt-4o-mini	79.09	84.0	81.36
gpt-4o	72.27	94.0	87.27
llama-3.1-8B-In	72.73	79.0	60.0
llama-3.3-70B-In	76.36	85.50	75.0
phi-4	88.0	75.91	84.09

(b) Syllogistic

Model	Incong.	Cong.	Nonsense
gpt-4o-mini	66.25	75.0	67.50
gpt-4o	78.75	83.13	73.12
llama-3.1-8B-In	66.87	60.62	41.25
llama-3.3-70B-In	71.25	71.25	70.63
phi-4	68.75	79.37	78.75

Table 6: Accuracy (%) per content type for Normative problems on each task (Zero-Shot). **Incong.** = Incongruent, **Cong.** = Congruent.

Content Effect. We observed content effects in both normative and epistemic problems. Table 6 presents the performance of the models on the normative problems of the different content types in each task. In the Deontic Logic task, most of the models except Phi-4 perform best on problems with congruent content, followed by those with nonsense and incongruent content. Phi-4 shows the highest performance on incongruent content, followed by nonsense and congruent content. In the Syllogistic task, problems with nonsense content tend to be more challenging; however, the performance gap narrows among the better-performing models.

6 Discussion

Our findings highlight several areas for improv-

Modal: Normative **Inference Pattern:** Mu-Mi
Gold Label: valid (entailment) **Content Type:** congruent

```
To determine whether the hypothesis follows from the premise, let's analyze both statements:

1. **Premise:** "You must take care of your health."
   - This statement implies an obligation or necessity to take care of your health. It suggests that taking care of your
     ↪ health is required.

2. **Hypothesis:** "You can choose to take care of your health."
   - This statement implies that taking care of your health is optional or a matter of choice. It suggests that you have
     ↪ the freedom to decide whether or not to take care of your health.

Now, let's compare the two:

- The premise indicates a requirement or obligation ("must"), while the hypothesis suggests an option or choice ("can
  ↪ choose").
- The premise does not imply that taking care of your health is optional; rather, it suggests it is necessary.
- Therefore, the hypothesis does not logically follow from the premise because the premise does not support the idea of
  ↪ having a choice in the matter.

Based on this analysis, the relationship between the premise and the hypothesis is one of non-entailment.

non-entailment
```

Figure 4: An example output of GPT-4o with a CoT prompt for reasoning from obligation to permission (Mu-Mi), which includes the expression "can choose to."

ing the logical consistency of LLMs' normative reasoning.

First, we observed that performance of the LLMs varies not only across different reasoning tasks but also within the same task, depending on the specific reasoning pattern. Our results on the Deontic Logic task reveal that the reasoning from obligation to permission (Mu-Mi) did not align with expected validity despite its simple form. This can lead to logical inconsistency, as LLMs tend to accept its logically equivalent contrapositive (NotMi-NotMu) as valid reasoning. Analysis of Chain-of-Thought outputs suggests that this inconsistency is often rooted in how permission is expressed in the problem. For instance, GPT-4o did not infer "You can choose to take care of your health" from "You must take care of your health", interpreting "can choose to" as an option rather than a statement of permission (see Figure 4). Our evaluation of syllogistic reasoning further indicates that negation affects the logical consistency of normative reasoning in LLMs, with reasoning patterns involving negation proving more difficult for the models. At the same time, the results from the Deontic Logic task suggest that the difficulty associated with negation is not solely due to its presence but may also depend on the complexity of the reasoning.

Second, we identified several biases and effects in normative reasoning of LLMs that resemble those observed in human reasoning. As with other kinds of reasoning in LLMs, normative reasoning is influenced by content effects, which affect the logical consistency of the models. We also observed domain specificity of normative reasoning. Our comparison of normative and epistemic reasoning indicates that whether the models perform better in the normative domain than in other domains depends on the specific task. This finding suggests that conclusions about model performance within a domain should take potential task-specific variations into account.

Throughout our experiments, we compared different prompting strategies, namely Zero-Shot, Few-Shot, and Chain-of-Thought, to examine their impact on the logical consistency of LLMs in normative reasoning. Our findings indicate that the effectiveness of these strategies varies across tasks and models. For Deontic Logic and Syllogistic reasoning tasks, Few-Shot prompting generally led to performance improvements, as expected. This can be attributed to the straightforward nature of leveraging similarities in syntactic structures within reasoning patterns. However, the observed improvements may result from superficial pattern matching rather than genuine reasoning. On the other hand, Chain-of-Thought prompting generally produced minimal improvement or even negative effects, a pattern also observed in epistemic reasoning. A common issue was the introduction of errors in intermediate reasoning steps, which led to incorrect final conclusions. These observations suggest that Chain-of-Thought does not necessarily enhance robustness in normative reasoning and may instead introduce additional points of failure.

7 Conclusion

In this paper, we systematically evaluated the normative reasoning capabilities of LLMs, examining both formal logical structures and cognitive influences. Our findings emphasize the significance of *formal* validity and invalidity in normative reasoning, extending beyond normative *content*. While models often align with valid reasoning patterns, they exhibit inconsistencies in specific inferences and cognitive biases similar to human reasoning. These results highlight challenges in maintaining formal consistency and underscore the need for further improvements to enhance the reliability of normative reasoning in LLMs.

Limitations

While our study provides a comprehensive analysis of normative reasoning in LLMs, several limitations should be acknowledged. First, the nature of normative reasoning and deontic logic remains an open research question. Our evaluation relies on a specific set of valid reasoning patterns based on the literature on logical and cognitive studies of normative reasoning, but alternative frameworks could yield different insights into LLM performance. Our study focuses on specific, controlled reasoning tasks rather than open-ended normative deliberation. While our dataset captures key logical patterns, real-world normative reasoning often involves additional complexities, such as contextual interpretation, ethical considerations, and pragmatic constraints. Future work could explore how LLMs perform in more applied settings in broader contexts.

Second, the LLMs are continuously evolving, and our findings may not generalize to future models with improved reasoning mechanisms. The performance of models is affected by updates in training data, architecture, and fine-tuning strategies. In addition, we compare closed and open-source models, but our experiments rely on API access for proprietary models, limiting our ability to analyze their internal mechanisms. Differences in fine-tuning, prompt engineering, and instruction-following abilities may also contribute to performance variations, making it challenging to isolate the effects of model architecture alone.

Acknowledgements

We thank the anonymous reviewers for their comments and suggestions. This work is partially supported by JST CREST Grant Number JPMJCR2114, Keio University Global Research Institute (KGRI) Challenge Grant, and JSPS Kakenhi Grant Numbers JP24K00004, JP21K00016, JP21H00467, JP23K20416, and JP21K18339.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Guilherme F.C.F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, 333:104145.

Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2023. Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 1–11.

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, Miami, Florida, USA. 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. 2025. Empowering LLMs with logical reasoning: A comprehensive survey. *arXiv preprint arXiv:2502.15652*.

Patricia W Cheng and Keith J Holyoak. 1985. Pragmatic reasoning schemas. *Cognitive psychology*, 17(4):391–416.

Agata Ciabattoni, John F. Horty, Marija Slavkovik, Leendert van der Torre, and Aleks Knoks. 2023. Normative reasoning for AI (Dagstuhl seminar 23151). *Dagstuhl Reports*, 13(4):1–23.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20)*, pages 3882–3890.

Leda Cosmides. 1989. The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, 31(3):187–276.

Leda Cosmides and John Tooby. 1992. Cognitive adaptations for social exchange. In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press.

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*.

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 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8425–8444.

Jonathan St.B. T. Evans, Stephen E. Newstead, and Ruth M. J. Byrne. 1993. *Human Reasoning: The Psychology of Deduction*. Psychology Press.

Laurence Fiddick. 2004. Domains of deontic reasoning: Resolving the discrepancy between the cognitive and moral reasoning literatures. *The Quarterly Journal of Experimental Psychology Section A*, 57(3):447–474.

D.M. Gabbay, J. Horty, and X. Parent. 2013. *Handbook of Deontic Logic and Normative Systems*. College Publications.

Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a dataset: A large negation benchmark to challenge large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8615, Singapore. Association for Computational Linguistics.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Jaakko Hintikka. 1962. *Knowledge and Belief*. Cornell University Press.

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.

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

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.

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.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Paul McNamara and Frederik Van De Putte. 2022. Deontic Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2022 edition. Metaphysics Research Lab, Stanford University.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. In *First Conference on Language Modeling*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(4):1–37.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

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*.

Rasmus Rendsvig, John Symons, and Yanjing Wang. 2024. Epistemic Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Summer 2024 edition. Metaphysics Research Lab, Stanford University.

Alf Ross. 1941. Imperatives and logic. *Theoria*, 7(1):53–71.

Spencer M. Seals and Valerie L. 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, Mexico City, Mexico. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4279–4303.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not

naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Georg Henrik von Wright. 1951. Deontic logic. *Mind*, 60(237):1–15.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

A Dataset Description

Epistemic Syllogism. All formal patterns of epistemic syllogisms examined in this work can be found in Table 7.

Label: Cat-MP P1: All B are certain to C P2: A is B C: A is certain to C	**Label**: Cat-MT P1: All B are certain to C P2: A is not certain to C C: A is not B	**Label**: Hyp-MP P1: If A is B, then it is certain that A is C P2: A is B C: It is certain that A is C	**Label**: Hyp-MT P1: If A is B, then it is certain that A is C P2: It is not certain that A is C C: A is not B
Label: Cat-AC P1: All B are certain to C P2: A is certain to C C: A is B	**Label**: Cat-DA P1: All B are certain to C P2: A is not B C: A is not certain to C	**Label**: Hyp-AC P1: If A is B, then it is certain that A is C P2: It is certain that A is C C: A is B	**Label**: Hyp-DA P1: If A is B, then it is certain that A is C P2: A is not B C: It is not certain that A is C

Table 7: 8 patterns of epistemic syllogisms. P1: Major Premise, P2: Minor Premise, C: Conclusion; Cat: Categorical, Hyp: Hypothetical; MP: Modus Ponens, MT: Modus Tollens, AC: Affirming the Consequent, DA: Denying the Antecedent. B and C represent predicates, and A represents a term. Those in blue are *valid* patterns, while those in red are *invalid* patterns.

Expressions of Modality. Table 8 and Table 9 present examples of natural language expressions for all types of normative and epistemic modalities (and negations) examined in this work.

Form	Example
Mu (*must* A)	*It is obligatory to* A*, It is mandatory to* A*, There is an obligation to* A*, You are required to* A*, You must* A
Mi (*might* A)	*It is permissible to* A*, It is acceptable to* A*, You are permitted to* A*, You are allowed to* A*, You can choose to* A
MuNot (*must not* A)	*It is obligatory not to* A*, It is mandatory not to* A*, There is an obligation not to* A*, You are required not to* A*, You must not* A
MiNot (*might not* A)	*It is permissible not to* A*, It is acceptable not to* A*, You are permitted not to* A*, You are allowed not to* A*, You can choose not to* A
NotMu (*not required to* A)	*It is not obligatory to* A*, It is not mandatory to* A*, There is no obligation to* A*, You are not required to* A*, It is not the case that you must* A
NotMi (*not permitted to* A)	*It is not permissible to* A*, It is not acceptable to* A*, You are not permitted to* A*, You are not allowed to* A*, You cannot choose to* A

Table 8: 6 types of normative modality (and negations) and examples of natural language expressions for each. Mu: obligation, Mi: permission, MuNot: obligation of negation, MiNot: permission of negation, NotMu: negations of obligation, NotMi: negations of permission.

Form	Example
Mu (S *must* P)	*It is certain that* S P*, It is necessarily the case that* S P*, It is known that* S P*, It is established that* S P*,* S *must have been* P
Mi (S *might* P)	*It is possible that* S P*,* S *might* P*,* S *might have been* P*, There's a chance* S P*, There is a possibility that* S P
MuNot (S *must not* P)	*It is certain that* S *not* P*, It is necessarily the case that* S *not* P*, It is known that* S *not* P*, It is established that* S *not* P*,* S *could not have been* P
MiNot (S *might not* P)	*It is possible that* S *not* P*,* S *might not* P*,* S *might not have been* P*, There's a chance* S *not* P*, There is a possibility that* S *not* P
NotMu (*not certain that* S P)	*It is not certain that* S P*, It is not necessarily the case that* S P*, It is not known that* S P*, It is not established that* S P*, It is not the case that* S *must have been* P
NotMi (*impossible that* S P)	*It is impossible that* S P*, It is not the case that* S *might have been* P*, There's no chance* S P*, There is no possibility that* S P

Table 9: 6 types of epistemic modality (and negations) and examples of natural language expressions for each. Mu: epistemic necessity, Mi: epistemic possibility, MuNot: epistemic necessity of negation, MiNot: epistemic possibility of negation, NotMu: negations of epistemic necessity, NotMi: negations of epistemic possibility.

B Prompt Templates and Examples

Tables 10 and 11 present examples of the prompts used for the Deontic Logic task and Syllogistic task, respectively.

Zero-Shot prompt example (Normative)

```
Determine whether the hypothesis follows from the premise(s).
- Answer 'entailment' if the hypothesis follows from the premise(s).
- Otherwise, answer 'non-entailment'.
Respond only with 'entailment' or 'non-entailment', and nothing else.

Premise: You are not required to attend the meeting.
Hypothesis: You are permitted not to attend the meeting.
Answer:
```

Few-Shot prompt example (Normative, abbreviated)

```
Determine whether the hypothesis follows from the premise(s).
- Answer 'entailment' if the hypothesis follows from the premise(s).
- Otherwise, answer 'non-entailment'.
Respond only with 'entailment' or 'non-entailment', and nothing else.

Premise: You are not required to finish homework by Friday.
Hypothesis: It is permissible not to finish homework by Friday.
Answer: entailment

[...]

Premise: It is obligatory to mail a letter.
Hypothesis: It is obligatory to mail a letter or to burn it.
Answer: non-entailment

Premise: You are not required to attend the meeting.
Hypothesis: You are permitted not to attend the meeting.
Answer:
```

CoT prompt example (Normative)

```
Determine whether the hypothesis follows from the premise(s).
- Answer 'entailment' if the hypothesis follows from the premise(s).
- Otherwise, answer 'non-entailment'.
Let's think step by step. Then output only one word, 'entailment' or 'non-entailment' on the last
    ↪ line, immediately after a line break.

Premise: You are not required to attend the meeting.
Hypothesis: You are permitted not to attend the meeting.
Answer:
```

Table 10: Prompt Examples for Deontic Logic Task.

Zero-Shot prompt example (Normative)

```
Determine whether the hypothesis follows from the premise(s).
- Answer 'entailment' if the hypothesis follows from the premise(s).
- Otherwise, answer 'non-entailment'.
Respond only with 'entailment' or 'non-entailment', and nothing else.

Premise: If you are a student, then it is obligatory to attend class. You are a student.
Hypothesis: It is obligatory to attend class.
Answer:
```

Few-Shot prompt example (Normative, abbreviated)

```
Determine whether the hypothesis follows from the premise(s).
- Answer 'entailment' if the hypothesis follows from the premise(s).
- Otherwise, answer 'non-entailment'.
Respond only with 'entailment' or 'non-entailment', and nothing else.

Premise: If the traffic light is red, then Taro must stop the car. The traffic light is red.
Hypothesis: Taro must stop the car.
Answer: entailment

[...]

Premise: All first-year students at this university must submit a paper. Bob is not a first-year
    ↪ student at this university.
Hypothesis: Bob is not required to submit a paper.
Answer: non-entailment

Premise: If you are a student, then it is obligatory to attend class. You are a student.
Hypothesis: It is obligatory to attend class.
Answer:
```

CoT prompt example (Normative)

```
Determine whether the hypothesis follows from the premise(s).
- Answer 'entailment' if the hypothesis follows from the premise(s).
- Otherwise, answer 'non-entailment'.
Let's think step by step. Then output only one word, 'entailment' or 'non-entailment' on the last
    ↪ line, immediately after a line break.

Premise: If you are a student, then it is obligatory to attend class. You are a student.
Hypothesis: It is obligatory to attend class.
Answer:
```

Table 11: Prompt Examples for Syllogistic Task.

C Supplemental Experimental Results

C.1 Deontic Logic Task

Examples of errors for each inference pattern are presented in Figures 5, 6, and 7. The complete results for the Deontic Logic task are reported in Tables 12 and 13.

Modal: Normative **Inference Pattern:** Mu-Mi
Gold Label: valid **Content Type:** congruent

P1: You must follow the rules of the game.
C: You can choose to follow the rules of the game.

Model (Few-shot)	Prediction
gpt-4o-mini	✗ invalid
gpt-4o	✗ invalid
llama-3.1-8b-In	✗ invalid
llama-3.3-70b-In	✗ invalid
phi-4	✗ invalid

Modal: Normative **Inference Pattern:** Mu-Mi
Gold Label: valid **Content Type:** congruent

P1: You must take care of your health.
C: You can choose to take care of your health.

Model (Few-shot)	Prediction
gpt-4o-mini	✗ invalid
gpt-4o	✗ invalid
llama-3.1-8b-In	✗ invalid
llama-3.3-70b-In	✓ valid
phi-4	✓ valid

Figure 5: Examples of errors for Mu-Mi.

Modal: `Normative` **Inference Pattern:** `FC-Or-Intro`
Gold Label: `invalid` **Content Type:** `congruent`

P1: You may eat cake.
C: You may eat cake or cookies.

Model (Zero-shot)	Prediction
gpt-4o-mini	✗ `valid`
gpt-4o	✗ `valid`
llama-3.1-8b-In	✓ `invalid`
llama-3.3-70b-In	✗ `valid`
phi-4	✓ `invalid`

Figure 6: An example of error for `FC-Or-Intro`.

Modal: `Normative` **Inference Pattern:** `Ross-Or-Intro`
Gold Label: `invalid` **Content Type:** `incongruent`

P1: You must pay your taxes.
C: You must pay your taxes or evade them.

Model (Zero-shot)	Prediction
gpt-4o-mini	✓ `invalid`
gpt-4o	✗ `valid`
llama-3.1-8b-In	✗ `valid`
llama-3.3-70b-In	✗ `valid`
phi-4	✗ `valid`

Figure 7: An example of error for `Ross-Or-Intro`.

(a) gold = *entailment*.

Model	NotMu-MiNot				NotMi-MuNot				MiNot-NotMu				Mu-Mi				NotMi-NotMu				FC-Or-Elim			
	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.
gpt-4o-mini	100.0	100.0	100.0	100.0	100.0	45.0	85.0	76.67	100.0	100.0	100.0	100.0	35.0	70.0	10.0	38.33	80.0	85.0	80.0	81.67	25.0	15.0	100.0	46.67
+ Few-Shot	100.0	95.0	100.0	98.33	95.0	35.0	35.0	55.0	100.0	100.0	100.0	100.0	80.0	95.0	80.0	85.0	100.0	95.0	95.0	96.67	55.0	35.0	35.0	41.67
+ CoT	100.0	90.0	100.0	96.67	65.0	35.0	70.0	56.67	100.0	95.0	100.0	98.33	30.0	10.0	15.0	18.33	90.0	100.0	80.0	90.0	60.0	20.0	90.0	56.67
gpt-4o	100.0	90.0	100.0	96.67	100.0	45.0	70.0	71.67	100.0	100.0	100.0	100.0	70.0	55.0	45.0	56.67	100.0	100.0	100.0	100.0	70.0	10.0	90.0	56.67
+ Few-Shot	100.0	90.0	100.0	96.67	100.0	85.0	85.0	90.0	100.0	100.0	100.0	100.0	80.0	85.0	80.0	81.67	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ CoT	100.0	90.0	100.0	96.67	60.0	0.0	35.0	31.67	95.0	90.0	100.0	95.0	80.0	75.0	80.0	78.33	80.0	85.0	70.0	78.33	80.0	10.0	100.0	63.33
llama-3.1-8B-In	100.0	100.0	100.0	100.0	100.0	80.0	100.0	93.33	100.0	100.0	95.0	98.33	60.0	30.0	25.0	38.33	85.0	80.0	95.0	86.67	45.0	25.0	10.0	26.67
+ Few-Shot	100.0	95.0	95.0	96.67	100.0	70.0	85.0	85.0	100.0	90.0	90.0	93.33	45.0	0.0	0.0	15.0	100.0	95.0	95.0	96.67	0.0	0.0	0.0	0.0
+ CoT	65.0	55.0	70.0	63.33	75.0	50.0	55.0	60.0	60.0	55.0	65.0	60.0	35.0	20.0	45.0	33.33	45.0	60.0	50.0	51.67	50.0	35.0	45.0	43.33
llama-3.3-70B-In	100.0	95.0	100.0	98.33	100.0	55.0	100.0	85.0	100.0	95.0	100.0	98.33	95.0	95.0	55.0	81.67	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0
+ Few-Shot	100.0	95.0	100.0	98.33	100.0	90.0	100.0	96.67	100.0	100.0	100.0	100.0	90.0	85.0	65.0	80.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ CoT	95.0	80.0	100.0	91.67	95.0	55.0	95.0	81.67	90.0	95.0	95.0	93.33	70.0	90.0	80.0	80.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
phi-4	100.0	80.0	100.0	93.33	80.0	30.0	90.0	66.67	75.0	75.0	80.0	76.67	85.0	75.0	45.0	68.33	95.0	95.0	90.0	93.33	100.0	100.0	100.0	100.0
+ Few-Shot	90.0	60.0	100.0	83.33	75.0	10.0	30.0	38.33	75.0	70.0	100.0	81.67	85.0	100.0	90.0	91.67	100.0	100.0	95.0	98.33	100.0	100.0	100.0	100.0
+ CoT	95.0	65.0	90.0	83.33	65.0	30.0	50.0	48.33	80.0	75.0	75.0	76.67	60.0	65.0	75.0	66.67	75.0	90.0	45.0	70.0	100.0	100.0	100.0	100.0

(b) gold = *non-entailment*.

Model	NotMu-NotMi				MiNot-MuNot				Mi-Mu				FC-Or-Intro				Ross-Or-Intro			
	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.
gpt-4o-mini	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	100.0	91.67	-	80.0	20.0	50.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	100.0	60.0	80.0
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	100.0	93.33	-	100.0	45.0	72.5
gpt-4o	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0	100.0	90.0	-	25.0	55.0	40.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	100.0	100.0	100.0
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	100.0	100.0	100.0
llama-3.1-8B-In	70.0	80.0	15.0	55.0	55.0	60.0	25.0	46.67	75.0	85.0	50.0	70.0	100.0	100.0	100.0	100.0	-	60.0	45.0	52.5
+ Few-Shot	90.0	95.0	60.0	81.67	65.0	90.0	60.0	71.67	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	65.0	60.0	62.5
+ CoT	70.0	60.0	55.0	61.67	20.0	55.0	25.0	33.33	30.0	60.0	70.0	53.33	65.0	65.0	90.0	73.33	-	25.0	30.0	27.5
llama-3.3-70B-In	100.0	100.0	100.0	100.0	60.0	80.0	70.0	70.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	20.0	0.0	10.0
+ Few-Shot	100.0	100.0	100.0	100.0	60.0	65.0	60.0	61.67	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	100.0	100.0	100.0
+ CoT	100.0	100.0	100.0	100.0	70.0	80.0	85.0	78.33	100.0	100.0	100.0	100.0	30.0	5.0	25.0	20.0	-	50.0	0.0	25.0
phi-4	100.0	100.0	100.0	100.0	60.0	80.0	60.0	66.67	100.0	100.0	100.0	100.0	85.0	35.0	90.0	70.0	-	65.0	70.0	67.5
+ Few-Shot	100.0	100.0	100.0	100.0	65.0	100.0	70.0	78.33	100.0	100.0	100.0	100.0	75.0	25.0	10.0	36.67	-	100.0	65.0	82.5
+ CoT	100.0	95.0	95.0	96.67	85.0	90.0	70.0	81.67	95.0	100.0	100.0	98.33	65.0	35.0	100.0	66.67	-	100.0	50.0	75.0

Table 12: Accuracy (%) for **Normative** Problems in Deontic Logic Task. **C** = Congruent, **I** = Incongruent, **N** = Nonsense, **Avg.** = Average.

(a) gold = *entailment*.

Model	NotMu-MiNot				NotMi-MuNot				MiNot-NotMu				Mu-Mi				NotMi-NotMu			
	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.
gpt-4o-mini	100.0	100.0	100.0	100.0	85.0	55.0	80.0	73.33	100.0	85.0	85.0	90.0	85.0	15.0	85.0	61.67	80.0	20.0	65.0	55.0
+ Few-Shot	100.0	100.0	100.0	100.0	45.0	55.0	80.0	60.0	100.0	80.0	100.0	93.33	90.0	35.0	100.0	75.0	90.0	20.0	50.0	53.33
+ CoT	100.0	100.0	100.0	100.0	80.0	85.0	85.0	83.33	100.0	90.0	100.0	96.67	35.0	25.0	65.0	41.67	55.0	15.0	35.0	35.0
gpt-4o	100.0	100.0	100.0	100.0	85.0	85.0	90.0	86.67	100.0	90.0	100.0	96.67	90.0	85.0	100.0	91.67	100.0	30.0	95.0	75.0
+ Few-Shot	100.0	90.0	100.0	96.67	90.0	100.0	85.0	91.67	100.0	80.0	100.0	93.33	100.0	75.0	100.0	91.67	80.0	30.0	85.0	65.0
+ CoT	100.0	100.0	100.0	100.0	85.0	95.0	85.0	88.33	100.0	75.0	95.0	90.0	85.0	90.0	100.0	91.67	70.0	25.0	95.0	63.33
llama-3.1-8B-In	100.0	100.0	100.0	100.0	100.0	90.0	100.0	96.67	100.0	100.0	100.0	100.0	35.0	50.0	35.0	40.0	100.0	65.0	100.0	88.33
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	80.0	95.0	91.67	100.0	100.0	90.0	96.67	80.0	65.0	85.0	76.67	95.0	65.0	85.0	81.67
+ CoT	55.0	75.0	60.0	63.33	75.0	60.0	95.0	76.67	55.0	35.0	60.0	50.0	5.0	0.0	5.0	3.33	55.0	15.0	70.0	46.67
llama-3.3-70B-In	100.0	100.0	100.0	100.0	95.0	70.0	85.0	83.33	100.0	95.0	100.0	98.33	100.0	90.0	100.0	96.67	80.0	20.0	80.0	60.0
+ Few-Shot	100.0	100.0	100.0	100.0	90.0	90.0	100.0	93.33	100.0	80.0	100.0	93.33	95.0	80.0	70.0	81.67	85.0	20.0	80.0	61.67
+ CoT	100.0	100.0	100.0	100.0	100.0	95.0	100.0	98.33	100.0	100.0	100.0	100.0	100.0	95.0	100.0	98.33	100.0	30.0	100.0	76.67
phi-4	100.0	100.0	100.0	100.0	100.0	95.0	90.0	95.0	100.0	95.0	95.0	96.67	100.0	80.0	100.0	93.33	90.0	25.0	85.0	66.67
+ Few-Shot	100.0	100.0	100.0	100.0	30.0	55.0	55.0	46.67	85.0	70.0	90.0	81.67	100.0	70.0	100.0	90.0	75.0	20.0	85.0	60.0
+ CoT	95.0	100.0	100.0	98.33	95.0	90.0	95.0	93.33	100.0	70.0	100.0	90.0	80.0	70.0	80.0	76.67	70.0	15.0	60.0	48.33

(b) gold = *non-entailment*.

Model	NotMu-NotMi				MiNot-MuNot				Mi-Mu			
	C	I	N	Avg.	C	I	N	Avg.	C	I	N	Avg.
gpt-4o-mini	95.0	100.0	100.0	98.33	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
gpt-4o	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
llama-3.1-8B-In	70.0	55.0	0.0	41.67	75.0	70.0	40.0	61.67	85.0	90.0	65.0	80.0
+ Few-Shot	30.0	95.0	20.0	48.33	100.0	90.0	85.0	91.67	100.0	100.0	100.0	100.0
+ CoT	95.0	100.0	100.0	98.33	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
llama-3.3-70B-In	85.0	100.0	100.0	95.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ CoT	75.0	100.0	100.0	91.67	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
phi-4	75.0	100.0	80.0	85.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 13: Accuracy (%) for **Epistemic** Problems in Deontic Logic Task. **C** = Congruent, **I** = Incongruent, **N** = Nonsense, **Avg.** = Average.

C.2 Syllogistic Task

Examples of errors for each inference pattern are presented in Figures 8, 9, 10, and 11. The complete results for the Syllogistic task are reported in Tables 14 and 15.

Modal: `Normative` **Inference Pattern:** `Cat-DA`
Gold Label: `invalid` **Content Type:** `congruent`

P1: All customers must pay for their purchases.
P2: A window shopper is not a customer.
C: There is no obligation for a window shopper to pay for their purchases.

Model (Few-shot)	Prediction
gpt-4o-mini	✗ `valid`
gpt-4o	✗ `valid`
llama-3.1-8b-In	✗ `valid`
llama-3.3-70b-In	✗ `valid`
phi-4	✗ `valid`

Modal: `Epistemic` **Inference Pattern:** `Cat-DA`
Gold Label: `invalid` **Content Type:** `incongruent`

P1: All flowers necessarily have petals.
P2: Sunflowers are not flowers.
C: Sunflowers do not necessarily have petals.

Model (Few-shot)	Prediction
gpt-4o-mini	✗ `valid`
gpt-4o	✗ `valid`
llama-3.1-8b-In	✗ `valid`
llama-3.3-70b-In	✗ `valid`
phi-4	✓ `invalid`

Figure 8: Examples of errors for `Cat-DA`.

Modal: Normative **Inference Pattern:** Hyp-DA
Gold Label: invalid **Content Type:** congruent

P1: If you want to drive a car, then it is obligatory to have a driver's license.
P2: You do not want to drive a car.
C: It is not obligatory to have a driver's license.

Model (Few-shot)	Prediction
gpt-4o-mini	✓ invalid
gpt-4o	✓ invalid
llama-3.1-8b-In	✗ valid
llama-3.3-70b-In	✗ valid
phi-4	✓ invalid

Modal: Epistemic **Inference Pattern:** Hyp-DA
Gold Label: invalid **Content Type:** congruent

P1: If the expedition locates the ancient ruins, then it is established that they will make a significant archaeological discovery.
P2: The expedition did not locate the ancient ruins.
C: It is not established that the expedition will make a significant archaeological discovery.

Model (Few-shot)	Prediction
gpt-4o-mini	✗ valid
gpt-4o	✓ invalid
llama-3.1-8b-In	✗ valid
llama-3.3-70b-In	✗ valid
phi-4	✓ invalid

Figure 9: Examples of errors for Hyp-DA.

Modal: Normative **Inference Pattern:** Cat-MT
Gold Label: valid **Content Type:** congruent

P1: All teachers must be qualified.
P2: It is not the case that she must be qualified.
C: She is not a teacher.

Model (Few-shot)	Prediction
gpt-4o-mini	✓ valid
gpt-4o	✓ valid
llama-3.1-8b-In	✗ invalid
llama-3.3-70b-In	✓ valid
phi-4	✗ invalid

Figure 10: An example of an error for Cat-MT.

Modal: Normative **Inference Pattern:** Hyp-MT
Gold Label: valid **Content Type:** congruent

P1: If you are invited to a wedding, then it is customary to bring a gift.
P2: It is not the case that you are required to bring a gift.
C: You are not invited to a wedding.

Model (Few-shot)	Prediction
gpt-4o-mini	✓ valid
gpt-4o	✓ valid
llama-3.1-8b-In	✗ invalid
llama-3.3-70b-In	✓ valid
phi-4	✗ invalid

Figure 11: An example of an error for Hyp-MT.

(a) gold = *entailment*.

Model	Cat-MP				Cat-MT				Hyp-MP				Hyp-MT			
	C	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**
gpt-4o-mini	100.0	100.0	100.0	100.0	90.0	20.0	25.0	45.0	100.0	100.0	100.0	100.0	50.0	30.0	100.0	60.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	65.0	45.0	70.0	100.0	100.0	100.0	100.0	70.0	30.0	100.0	66.67
+ CoT	100.0	100.0	100.0	100.0	85.0	35.0	25.0	48.33	100.0	100.0	100.0	100.0	50.0	50.0	100.0	66.67
gpt-4o	100.0	100.0	100.0	100.0	90.0	85.0	25.0	66.67	100.0	100.0	100.0	100.0	85.0	90.0	100.0	91.67
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	80.0	93.33	100.0	100.0	100.0	100.0	100.0	60.0	100.0	86.67
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	25.0	75.0	100.0	100.0	100.0	100.0	60.0	70.0	100.0	76.67
llama-3.1-8B-In	100.0	100.0	100.0	100.0	90.0	95.0	25.0	70.0	100.0	95.0	100.0	98.33	80.0	50.0	100.0	76.67
+ Few-Shot	100.0	100.0	60.0	86.67	65.0	85.0	15.0	55.0	100.0	100.0	100.0	100.0	70.0	95.0	100.0	88.33
+ CoT	100.0	95.0	90.0	95.0	65.0	35.0	15.0	38.33	100.0	100.0	100.0	100.0	25.0	25.0	80.0	43.33
llama-3.3-70B-In	100.0	100.0	100.0	100.0	90.0	60.0	20.0	56.67	100.0	100.0	100.0	100.0	50.0	55.0	100.0	68.33
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	30.0	76.67	100.0	100.0	100.0	100.0	95.0	55.0	100.0	83.33
+ CoT	100.0	95.0	100.0	98.33	85.0	90.0	30.0	68.33	100.0	100.0	100.0	100.0	50.0	50.0	95.0	65.0
phi-4	100.0	100.0	100.0	100.0	100.0	25.0	30.0	51.67	100.0	100.0	100.0	100.0	45.0	35.0	100.0	60.0
+ Few-Shot	75.0	95.0	20.0	63.33	90.0	10.0	5.0	35.0	65.0	95.0	100.0	86.67	30.0	10.0	85.0	41.67
+ CoT	100.0	100.0	95.0	98.33	100.0	90.0	25.0	71.67	100.0	100.0	100.0	100.0	60.0	75.0	100.0	78.33

(b) gold = *non-entailment*.

Model	Cat-AC				Cat-DA				Hyp-AC				Hyp-DA			
	C	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**
gpt-4o-mini	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	55.0	75.0	15.0	48.33	55.0	75.0	15.0	48.33
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	85.0	90.0	100.0	91.67	85.0	90.0	100.0	91.67
gpt-4o	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	85.0	60.0	73.33	75.0	85.0	60.0	73.33
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.0	100.0	98.33	100.0	95.0	100.0	98.33
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
llama-3.1-8B-In	35.0	95.0	5.0	45.0	35.0	95.0	5.0	45.0	25.0	30.0	0.0	18.33	25.0	30.0	0.0	18.33
+ Few-Shot	0.0	100.0	0.0	33.33	0.0	100.0	0.0	33.33	0.0	5.0	0.0	1.67	0.0	5.0	0.0	1.67
+ CoT	50.0	100.0	50.0	66.67	50.0	100.0	50.0	66.67	50.0	55.0	10.0	38.33	50.0	55.0	10.0	38.33
llama-3.3-70B-In	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	25.0	50.0	45.0	40.0	25.0	50.0	45.0	40.0
+ Few-Shot	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	50.0	55.0	100.0	68.33	50.0	55.0	100.0	68.33
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	40.0	30.0	0.0	23.33	40.0	30.0	0.0	23.33
phi-4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0	75.0	100.0	81.67	70.0	75.0	100.0	81.67
+ Few-Shot	80.0	20.0	80.0	60.0	80.0	20.0	80.0	60.0	100.0	95.0	100.0	98.33	100.0	95.0	100.0	98.33
+ CoT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	85.0	95.0	86.67	80.0	85.0	95.0	86.67

Table 14: Accuracy (%) for **Normative** Problems in Syllogistic Task. **C** = Congruent, **I** = Incongruent, **N** = Nonsense, **Avg.** = Average.

(a) gold = *entailment*.

Model	Cat-MP				Cat-MT				Hyp-MP				Hyp-MT			
	C	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**
gpt-4o-mini	100.0	90.0	100.0	96.67	70.0	40.0	100.0	70.0	100.0	75.0	100.0	91.67	10.0	60.0	10.0	26.67
+ Few-Shot	100.0	75.0	100.0	91.67	80.0	65.0	95.0	80.0	100.0	70.0	100.0	90.0	10.0	30.0	5.0	15.0
+ CoT	100.0	100.0	100.0	100.0	90.0	45.0	85.0	73.33	100.0	85.0	100.0	95.0	5.0	30.0	0.0	11.67
gpt-4o	100.0	100.0	100.0	100.0	95.0	100.0	100.0	98.33	100.0	95.0	100.0	98.33	25.0	80.0	5.0	36.67
+ Few-Shot	100.0	100.0	100.0	100.0	90.0	95.0	90.0	91.67	100.0	95.0	100.0	98.33	40.0	15.0	45.0	33.33
+ CoT	100.0	100.0	95.0	98.33	75.0	80.0	85.0	80.0	100.0	85.0	100.0	95.0	0.0	25.0	0.0	8.33
llama-3.1-8B-In	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	100.0	91.67	100.0	85.0	85.0	90.0
+ Few-Shot	100.0	95.0	100.0	98.33	100.0	90.0	100.0	96.67	100.0	35.0	100.0	78.33	85.0	90.0	90.0	88.33
+ CoT	100.0	100.0	95.0	98.33	70.0	40.0	50.0	53.33	100.0	70.0	100.0	90.0	35.0	50.0	25.0	36.67
llama-3.3-70B-In	100.0	95.0	100.0	98.33	95.0	45.0	100.0	80.0	100.0	80.0	100.0	93.33	0.0	15.0	5.0	6.67
+ Few-Shot	100.0	95.0	100.0	98.33	95.0	45.0	100.0	80.0	100.0	80.0	100.0	93.33	0.0	15.0	5.0	6.67
+ CoT	100.0	95.0	100.0	98.33	95.0	45.0	100.0	80.0	100.0	80.0	100.0	93.33	0.0	15.0	5.0	6.67
phi-4	100.0	95.0	100.0	98.33	100.0	60.0	100.0	86.67	100.0	95.0	100.0	98.33	5.0	70.0	15.0	30.0
+ Few-Shot	100.0	90.0	100.0	96.67	100.0	70.0	100.0	90.0	100.0	50.0	100.0	83.33	0.0	10.0	0.0	3.33
+ CoT	100.0	100.0	100.0	100.0	100.0	85.0	100.0	95.0	100.0	100.0	100.0	100.0	20.0	50.0	5.0	25.0

(b) gold = *non-entailment*.

Model	Cat-AC				Cat-DA				Hyp-AC				Hyp-DA			
	C	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**	**C**	**I**	**N**	**Avg.**
gpt-4o-mini	70.0	75.0	100.0	81.67	70.0	75.0	100.0	81.67	5.0	10.0	20.0	11.67	5.0	10.0	20.0	11.67
+ Few-Shot	90.0	100.0	95.0	95.0	90.0	100.0	95.0	95.0	90.0	75.0	100.0	88.33	90.0	75.0	100.0	88.33
+ CoT	100.0	90.0	100.0	96.67	100.0	90.0	100.0	96.67	40.0	90.0	70.0	66.67	40.0	90.0	70.0	66.67
gpt-4o	95.0	55.0	85.0	78.33	95.0	55.0	85.0	78.33	5.0	35.0	50.0	30.0	5.0	35.0	50.0	30.0
+ Few-Shot	100.0	95.0	100.0	98.33	100.0	95.0	100.0	98.33	95.0	95.0	95.0	95.0	95.0	95.0	95.0	95.0
+ CoT	100.0	95.0	100.0	98.33	100.0	95.0	100.0	98.33	80.0	75.0	100.0	85.0	80.0	75.0	100.0	85.0
llama-3.1-8B-In	0.0	45.0	45.0	30.0	0.0	45.0	45.0	30.0	5.0	5.0	0.0	3.33	5.0	5.0	0.0	3.33
+ Few-Shot	0.0	40.0	15.0	18.33	0.0	40.0	15.0	18.33	0.0	5.0	0.0	1.67	0.0	5.0	0.0	1.67
+ CoT	15.0	70.0	50.0	45.0	15.0	70.0	50.0	45.0	20.0	30.0	40.0	30.0	20.0	30.0	40.0	30.0
llama-3.3-70B-In	70.0	85.0	95.0	83.33	70.0	85.0	95.0	83.33	5.0	20.0	30.0	18.33	5.0	20.0	30.0	18.33
+ Few-Shot	70.0	85.0	95.0	83.33	70.0	85.0	95.0	83.33	5.0	20.0	30.0	18.33	5.0	20.0	30.0	18.33
+ CoT	70.0	85.0	95.0	83.33	70.0	85.0	95.0	83.33	5.0	20.0	30.0	18.33	5.0	20.0	30.0	18.33
phi-4	95.0	75.0	100.0	90.0	95.0	75.0	100.0	90.0	30.0	90.0	35.0	51.67	30.0	90.0	35.0	51.67
+ Few-Shot	95.0	85.0	100.0	93.33	95.0	85.0	100.0	93.33	75.0	95.0	50.0	73.33	75.0	95.0	50.0	73.33
+ CoT	45.0	50.0	80.0	58.33	45.0	50.0	80.0	58.33	55.0	60.0	50.0	55.0	55.0	60.0	50.0	55.0

Table 15: Accuracy (%) for **Epistemic** Problems in Syllogistic Task. **C** = Congruent, **I** = Incongruent, **N** = Nonsense, **Avg.** = Average.