A Study on the Reasoning Ability of LLMs Using Negation Principles in Chinese Sentences

Ran Li[1],Lingxiang Fan[1], Ze Gan[1], Zhaoyang Gao[1], Lifei Wang[1], Renjia Xiao[1], Zhe Yu[2], Guiyun Zhao[1], Zhe Lin[1]

[1]Department of Philosophy, Xiamen University
[2]Institute of Logic and Cognition, Department of Philosophy, Sun Yat-sen University
yuzh28@mail.sysu.edu.cn, pennyshaq@163.com

Abstract

The reasoning abilities of large language models (LLMs) have attracted growing interest in both AI and Logic. In this paper, we examine the extent to which fourteen LLMs (eleven Chinese LLMs and three English LLMs) can understand the logical properties of negation in inferential contexts in Chinese. We focus on fifteen well-studied properties of negation in classical propositional logic, such as the inference: "If it rains, then the ground gets wet. Therefore: If the ground is not wet, then it did not rain." These principles are fundamental to deductive reasoning and serve as building blocks for various non-classical logics. The study of different forms of negation has long been central to philosophical logic, and assessing how LLMs handle negation-based inferences provides a more fine-grained evaluation of their reasoning capabilities, as well as a comparison with human logical competence.

Overall, the LLMs we tested perform reasonably well on negation reasoning in Chinese, but they show a clear resistance to certain negation properties: the models often fail to accept these rules and perform poorly on inferences that rely on them. Analysis by specific negation patterns reveals substantial differences: some forms of negation consistently pose greater difficulty, and models sometimes produce answers that directly contradict logical expectations. Zero-shot chain-of-thought prompting can improve consistency to some extent, but the degree of improvement varies across models and properties and is not consistently large. Under robustness tests, even high-performing models can experience notable declines. In addition, we observe that some models are more prone to adopt non-classical negation patterns, such as Intuitionistic negation or minimal negation rather than classical negation in certain cases.

Introduction

One of the most fundamental aspects of reasoning is the ability to derive conclusions from given premises, a cognitive capacity that lies at the core of human intelligence (33; 8; 28). While large language models (LLMs) have demonstrated performance approaching human capabilities in a wide range of tasks, logical reasoning remains a substantial challenge. Considerable research has therefore focused on enhancing the reasoning abilities of LLMs; for a comprehensive overview, see the recent survey by Cheng et al. (3).

Purely logical reasoning critically depends on understanding and correctly applying the properties of logical connectives. In this work, we focus on one of the most essential and foundational connectives: negation. The logical properties of negation have long attracted attention from philosophers, logicians, mathematicians, and computer scientists, and have given rise to numerous logical systems characterized by distinct negation properties (7).

Figure 1 summarizes the performance of several prominent LLMs on tasks requiring reasoning with basic negation properties. While some of the larger models exhibit strong performance, systematic errors remain evident. Note that in Figure 1, the first panel presents the evaluation results obtained under a temperature setting of 0, whereas the second panel corresponds to experiments conducted with a temperature setting of 1. A natural question therefore arises: to what extent do LLMs truly grasp negation reasoning, and does their understanding of negation align with human-like logical competence? In this study, we address this question from a logician's perspective by examining LLMs' mastery of fifteen fundamental negation inference patterns. For example, one such pattern investigates whether "If p then not q" entails "If q then not p." We constructed numerous natural language examples instantiating each pattern and evaluated fourteen different LLMs on their reasoning performance. For each pattern, we construct numerous natural-language instantiations such as:

> "If it cannot be proven that no proof exists for this problem, does it follow that a proof must exist? Answer only yes' or no'."

Models were prompted with fifteen distinct negation inference patterns, and accuracy was systematically recorded. All experiments were conducted under zero-shot conditions, and additional tests were performed using chain-of-thought prompting to assess the effect of intermediate reasoning steps (Figure 2). Furthermore, models exhibiting strong performance were subjected to robustness testing to evaluate the stability of their negation reasoning under varied conditions. Our experimental results reveal the following key findings.

Our experimental findings can be summarized as follows. First, all evaluated LLMs exhibit fundamental errors in inferential reasoning involving negation. Even the strongest models show substantially elevated error rates on certain

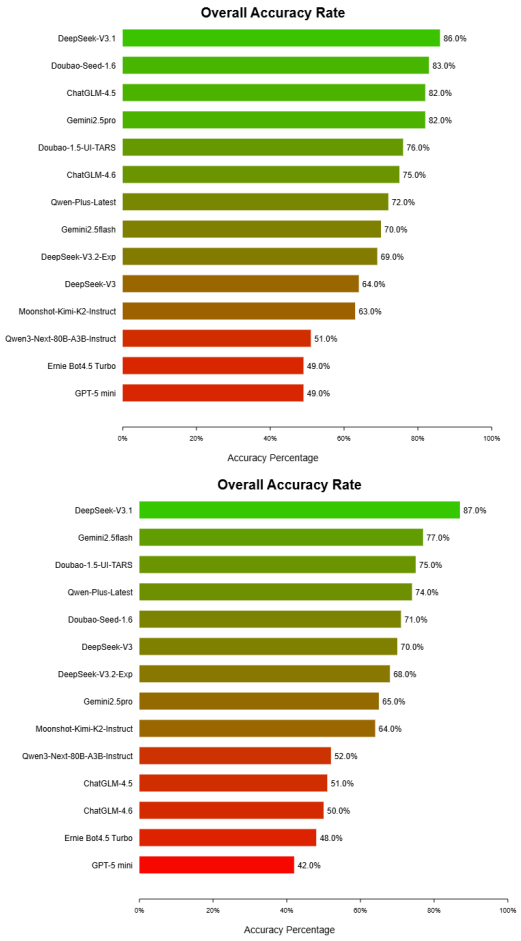


Figure 1: Overall Accuracy Rate

negation inference patterns. Second, while chain-of-thought prompting provides modest improvements, its overall effect remains limited, indicating that performance deficits are not merely due to insufficient intermediate reasoning steps. Third, different negation patterns yield markedly heterogeneous accuracy within each model.

Moreover, we observe a systematic bias in inferential behavior: LLMs tend to perform best on inference patterns validated in minimal, subminimal, and intuitionistic logics. Their accuracy on these non-classical forms of negation generally exceeds performance under classical and other non-classical negational frameworks. This lends empirical support to the hypothesis that LLMs implicitly approximate a weaker, non-classical conception of negation.

Finally, when subjected to robustness evaluations based on standard structural proof rules—including weakening, contraction, commutativity, and associativity—models that previously performed well show sharp declines in accuracy. This indicates that the negation-based reasoning of current LLMs lacks stability under routine variations in linguistic form and logical structure.This work advances evaluation methodology by systematically targeting inferential properties of negation across classical and non-classical logics. In doing so, it extends existing benchmarks, foregrounds the practical relevance of non-classical negation, and provides a principled basis for future assessments of non-classical reasoning in LLMs.

In this study, we focus on testing negation-based reasoning patterns expressed in Chinese for the following reasons: First most of the models we evaluate are large language models primarily trained in Chinese contexts, so using Chinese allows us to better assess their reasoning capabilities. Second our focus is on negation reasoning, and Chinese offers a wider variety of linguistic constructions for expressing negation. Moreover, the comparatively flexible syntactic structure of Chinese reduces the likelihood that models can rely on memorization or pattern matching rather than genuine reasoning during the evaluation.

Logical Background and Literature Review

A central concern in logic is the characterization of logical inference and its associated notion of logical consequence. Following the classical Tarskian account, an inference is logically valid if and only if it is impossible for its premises to be true while its conclusion is false, solely in virtue of the meanings of logical expressions—such as the sentential connectives "and," "or," "not," and "if" (33). Logical validity is thus independent of the specific subject matter represented by the non-logical vocabulary. For instance, from a conjunction "P and Q," the inference to "P" is valid regardless of what propositions P and Q denote.

By contrast, inferences relying on extra-logical lexical meaning—e.g., "Socrates is a father; therefore, Socrates has a child"—are not regarded as purely logical, since their correctness depends on empirical semantic content rather than on logical form. Hence, logical reasoning fundamentally requires understanding and correctly deploying logical operators.

Among these operators, negation plays a particularly crucial role. A paradigmatic example is the inference form known as *modus tollens*: from "If P, then Q" and "not Q," one may validly infer "not P." This rule relies on the classical principle of contraposition, which preserves truth across logically equivalent formulations. Conversely, common fallacies such as *denying the antecedent* illustrate how a misunderstanding of negation leads to invalid inferences: from "If P, then Q," one cannot infer "not P, therefore not Q," nor can one infer "not P, therefore Q."

In this study, we examine fifteen fundamental logical properties of negation (see Table 1), each of which is central to understanding and applying deductive reasoning correctly. However, classical logic does not capture all forms of consequence that arise naturally in mathematical and practical reasoning. Intuitionistic logic, for instance, reflects a constructive conception of inference, where a statement is accepted only when its truth can be established by proof (5). Likewise, many-valued logics such as Łukasiewicz's system represent reasoning under genuine indeterminacy by allowing propositions to take intermediate truth values, thereby modeling inference in contexts where classical bivalence does not hold (21). These non-classical logics show that alternative consequence relations emerge not from rejecting

classical principles but from analyzing different structural features of real reasoning. In all such frameworks, the behavior of negation is pivotal: the acceptance or rejection of specific negational properties determines the expressive power and inferential profile of each logic. The following discussion reviews the connections between key non-classical negations and the fundamental negation properties we have identified.

Turning to intuitionistic negation, Heyting's intuitionistic logic—developed to formalize Brouwer's constructivist programme (13)—defines negation by the clause $\neg\varphi := (\varphi \to \bot)$, whence the law of excluded middle $\vdash \varphi \textit{ or } \neg\varphi$ classical double negation $\neg\neg\varphi \vdash \varphi$ do not obtain intuitionistically. Moreover the distributive principle $\neg(\varphi \textit{ and } \psi) \vdash (\neg\varphi \textit{ or } \neg\psi)$ fails to hold in general. Nevertheless, the principle of absurdity (ex falso) persists, so that $\varphi \textit{ and } \neg\varphi \vdash \psi$ remains valid for arbitrary ψ.

A paradigmatic example frequently cited in the literature concerns undecided mathematical propositions such as the Goldbach conjecture: intuitionistically, one cannot assert "Goldbach is true *or* Goldbach is false" without either a constructive proof of the conjecture or a construction showing that any purported proof yields a contradiction; and to assert ¬Goldbach requires precisely such a construction leading from any proof of Goldbach to $\bot$. This constructive reading of negation—as a method of transforming any proof of φ into absurdity—has played a central role in (5; 29; 35).

Ortho negation extends intuitionistic negation by validating double negation ($\varphi \vdash \neg\neg\varphi$) but is defined on non-distributive lattices, such as ortholattices. As a result, both the law of excluded middle ($\varphi \textit{ or } \neg\varphi$) and some classical distributive laws fail. For instance, in general, $\neg(\varphi \textit{ and } (\psi \textit{ or } \chi)) \nvdash (\neg(\varphi \textit{ and } \psi)) \textit{ and } (\neg(\varphi \textit{ and } \chi))$. This failure reflects the non-distributive structure underlying ortho negation.

An intuitive example comes from quantum logic (7): let φ, ψ, χ represent spin properties along different axes. Ortho negation corresponds to the orthogonal complement of a subspace. While double negation holds ($\varphi \vdash (\varphi^{\perp})^{\perp}$), distributive identities fail because the orthocomplement of an intersection with a union does not entail the intersection of orthocomplements.

Minimal logic is obtained from Heyting's intuitionistic logic by removing the Duns Scotus law $\varphi \to (\neg\varphi \to \psi)$, thereby invalidating explosion while retaining the constructive principles $\varphi \vdash \neg\neg\varphi$ and, if $\varphi \vdash \neg\psi$, then $\psi \vdash \neg\varphi$. The concept of minimal negation was introduced by Kolmogorov (19) and systematically developed by Johansson (15). To illustrate the failure of absurdity, Johansson notes that although both "This object is red" and "This object is not red" may be derivable, minimal logic does not license inferring an unrelated statement such as "$2 + 2 = 5$"; contradictions do not trivialize reasoning. Minimal negation thus functions as constructive refutation rather than a truth-functional operator, allowing inconsistency without collapse into logical triviality.

Subminimal negation weakens minimal negation by rejecting both constructive double negation ($\varphi \nvdash \neg\neg\varphi$) and the De Morgan distributivity laws, while retaining only contraposition: from $\varphi \vdash \psi$ one may infer $\neg\psi \vdash \neg\varphi$. This system was first axiomatized by Hazen (12), who interpreted negation epistemically as *counterevidence* rather than classical falsity. Under this view, $\neg(\varphi \vee \psi)$ states merely that there is evidence against the disjunction as a whole, without providing counterevidence for either disjunct.

For example, let φ denote "it will rain tomorrow" and ψ denote "it will snow tomorrow." Forecasting data may indicate that there is *no basis to expect precipitation of any kind*, yielding $\neg(\varphi \vee \psi)$, while still offering no specific reason to accept either $\neg\varphi$ or $\neg\psi$ individually. Thus $\neg(\varphi \vee \psi) \nvdash \neg\varphi \wedge \neg\psi$, demonstrating that subminimal negation invalidates the De Morgan law for disjunction and captures a weak, evidence-based notion of negation. Došen's N-systems further show how one may partially restore structure by validating the distributivity of negation over conjunction, while still rejecting constructive double negation (4).

De Morgan negation, also called quasi-Boolean negation, extends minimal negation by adding the classical double-negation axiom $\neg\neg\varphi \vdash \varphi$. It validates constructive contraposition—if $\varphi \vdash \neg\psi$, then $\psi \vdash \neg\varphi$—and distributes over conjunction: $\neg(\varphi \wedge \psi) \vdash (\neg\varphi \vee \neg\psi)$. De Morgan negation has been widely studied in both algebra and logic, and its structure over distributive lattices is known as a De Morgan algebra. De Morgan algebras, or quasi-Boolean algebras, are introduced by Białynicki-Birula and Rasiowa (1), and further studied by Moisil and Kalman (23; 16).

A concrete example for De morgan negation comes from rough sets (27): a property $P \subseteq U$ is represented as a pair $(\underline{P}, \overline{P})$, where $\underline{P}$ contains objects that definitely satisfy P and $\overline{P}$ those that possibly satisfy P. The De Morgan negation of P is $(U \setminus \overline{P}, U \setminus \underline{P})$, capturing objects that definitely or possibly do not satisfy P. This pair-based representation satisfies double negation, distributes over conjunction, and prevents explosion, providing an intuitive, constructive model of De Morgan negation. These properties of negation are also relevant to a wide range of non-classical logics. For instance, in paraconsistent logics, the principle of explosion (from $\varphi \wedge \neg\varphi$ infer ψ) does not hold (30). In Kleene's three-valued logic, negation distributes over conjunction and disjunction but neither the law of excluded middle nor explosion is valid; nonetheless, it satisfies the inference from $\neg\psi \wedge \psi$ to $\neg\psi \vee \psi$ (17). Other forms of negation, including Nelson negation and fuzzy negation, also relate to the properties we discuss here, though we do not treat them exhaustively in this work (25; 9).

Overall, these examples illustrate that the logical properties of negation we have identified have a long-standing presence and broad relevance in both classical and non-classical logic. They have attracted the attention of logicians, philosophers, linguists, and mathematicians alike, reflecting their foundational role in the study of reasoning and formal semantics.

Research on the logical reasoning capabilities of large language models (LLMs) has become an increasingly important area of study. Numerous works have highlighted the limitations of LLM reasoning abilities, including (22; 24; 11). Studies focusing specifically on the pure logical reasoning capacity of LLMs have primarily explored this

through the development of formal benchmarks and evaluation criteria, as exemplified by a series of datasets created in recent years (**?** 34; 10; 31**?**). Notably, research on non-classical reasoning in LLMs, such as (20), has experimentally demonstrated that while LLMs exhibit certain non-monotonic reasoning patterns when handling generalizations and exceptions, they fail to maintain stable belief states when additional supporting or irrelevant evidence is introduced, and therefore cannot yet be considered human-level non-monotonic reasoners. Similarly, (14) has shown experimentally that LLMs display significant limitations when reasoning about conditionals and modals.

In the present study, we extend the work of (14) by evaluating LLM reasoning in the context of non-classical negations. We adopt a methodology similar to theirs but place greater emphasis on the fundamental properties of logical inference, focusing on single-step reasoning. This approach allows for a more precise analysis of the sources of errors in LLM reasoning under different negation properties. Consistent with prior work, our task format and evaluation method are similar to those used in the natural language inference (NLI) paradigm (2; 37; 26). However, our study differs in several key aspects. To better isolate pure reasoning ability, we design tasks in Chinese, a language with less overt syntactic cues than English. We concentrate specifically on reasoning involving various properties of non-classical negation. In addition, we develop robustness tests based on common fundamental logical properties of premises, including associativity, commutativity, contraction, and weakening. To the best of our knowledge, no prior work has systematically combined non-classical negation reasoning with such fundamental logical property testing in LLMs.

Experiments

We evaluated the logical inference capabilities of the fourteen LLMs listed in Figure 1, with detailed model specifications provided in Appendix A. All models were accessed via their respective APIs, and the full experimental dataset is publicly available at https://github.com/lianlian771/shiyuedidemougelunwen4.git.

For our experiments, we constructed a comprehensive question bank encompassing the fifteen negation inference patterns listed in the following Table 1, enabling a detailed evaluation of LLMs' reasoning capabilities. For each pattern, to mitigate potential confounds arising from memorization or pattern recognition by the models, we manually designed 32 instances per pattern, yielding a total of 475 questions (see Appendix B for representative examples). In total, we collected 104,172 raw responses. To prevent models from relying on world knowledge rather than the formal logical structure of the inference, we also incorporated novel, fictitious predicates within the questions. For instance, given "If Xincat arrived, then Beidog arrived," the LLM was required, based on the relevant negation principle, to infer "If Beidog did not arrive, then Xincat did not arrive."

For each problem and each LLM, instances were presented under both temperature settings of 0 and 1, using a zero-shot chain-of-thought prompting paradigm ((18); see Appendix C). Under the temperature-1 condition, each instance was queried three times per LLM to ensure response reliability and mitigate stochastic variability. Additionally, we performed robustness evaluations on the highest-performing configuration of each model. In these assessments, the premises were systematically manipulated by: concept substitution, redundancy and negation, division and recomposition and irrelevant interference. All experimental items and corresponding responses were manually reviewed and aligned by researchers with graduate-level training in logic, thereby ensuring rigorous verification of both reasoning correctness and consistency.

Results

The aggregate results of our reasoning evaluations are presented in Tables 2 and 3 (Appendix D). Table 2 reports overall performance under a temperature setting of 0, while Table 3 corresponds to a temperature of 1. The results indicate substantial variability across different large language models (LLMs). High-performing models, such as DeepSeek v3.1, Doubao-Seed 1.6 and Gemini 2.5 Pro, achieve accuracy approaching 87%, whereas lower-performing models exhibit accuracy near 50%. Although differences in temperature settings appear to influence model outputs, the effect is not statistically significant.

Figure 2 (Appendix D) illustrates that chain-of-thought (CoT) prompting yields a measurable improvement in reasoning performance across all evaluated LLMs. Notably, the magnitude of improvement is relatively uniform across models with differing baseline performance, with overall accuracy gains generally remaining below 5%. Nevertheless, for certain LLMs and specific negation inference patterns, error rates remain substantial even with CoT prompting. These findings corroborate prior evidence that CoT can elicit and enhance reasoning in LLMs (36), while also highlighting that its efficacy is not universal across all types of inference. Aggregate results following CoT prompting are reported in Tables 4 and 5. Based on Tables 2-5 (Appendix D), we subsequently highlight several particularly interesting observations pertaining to specific negation inference patterns.

In our investigation, we focused on several negation-related inference patterns where classical and non-classical logics diverge. Our objective was to assess whether large language models (LLMs) correctly interpret the intended meaning of negation and the associated logical principles. This examination yielded several notable findings. First, consider the distinction between classical double negation and constructive double negation. Logical theory shows that these principles do not coincide across systems: for example, in intuitionistic logic, constructive double negation is valid, whereas classical double negation is not. A familiar illustration is the inference from "It cannot be proven that this person is innocent" to "This person is guilty." While this inference is Boolean-valid in classical logic, it clashes with human intuitive reasoning. Our results indicate that LLMs behave similarly to humans: the vast majority correctly reject such inferences and answer no. (see Figure 3 (Appendix E))

In sharp contrast, evaluating constructive double negation, most LLMs again tend to answer yes to valid an inference

Valid Inferences	**Examples**
$\neg\neg\alpha \rightarrow \alpha$	If it is not the case that he cannot testify for you, does it mean that he can testify for you?
$\alpha \rightarrow \neg\neg\alpha$	If Xiao Ming is really eating pizza, does it mean that it is not the case that Xiao Ming is not eating pizza?
$\alpha \wedge \neg\alpha \rightarrow \beta$	A person simultaneously believes that traditional Chinese medicine cannot treat cancer and that traditional Chinese medicine can treat cancer. Does it mean that he believes that the circle is square?
$\alpha \wedge \neg\alpha \rightarrow \bot$	A philosopher writes: "There is no truth," and insists that this statement is true. Does it mean that he considers falsehood to be true?
$\beta \rightarrow (\alpha \vee \neg\alpha)$	The circle is square, so either traditional Chinese medicine cannot treat cancer or it can treat cancer. Is the above reasoning correct?
$\top \rightarrow (\alpha \vee \neg\alpha)$	Because one plus one either equals 2 or does not equal 2, so "either dark matter exists or dark matter does not exist."
If $(\alpha \rightarrow \beta)$, then $(\neg\beta \rightarrow \neg\alpha)$	The meteorologist said: "If the typhoon lands, there will be strong winds and heavy rain." Now there are no strong winds and heavy rain, so the typhoon did not land.
If $(\alpha \rightarrow \neg\beta)$, then $(\beta \rightarrow \neg\alpha)$	If we do bad things, then people will not harm us; therefore, if people harm us, we did not do bad things.
$\neg(\alpha \wedge \beta) \rightarrow (\neg\alpha \vee \neg\beta)$	We all know that you cannot have both. Does it mean that you must not have one or not have the other?
$(\neg\alpha \vee \neg\beta) \rightarrow \neg(\alpha \wedge \beta)$	If tonight the cinema is not closed or the amusement park is not closed, can it be concluded that they are not both closed?
$(\neg\alpha \wedge \neg\beta) \rightarrow \neg(\alpha \vee \beta)$	Sin is neither punished nor rewarded. Someone therefore infers: "Sin will not be punished or rewarded."
$\neg(\alpha \vee \beta) \rightarrow (\neg\alpha \wedge \neg\beta)$	Proposition P is not true or false; does it mean that proposition P is neither true nor false?
If $(\alpha \wedge \beta) \rightarrow r$, then $(\alpha \wedge \neg r) \rightarrow \neg\beta$	If goodness is an objective entity and people can know about goodness, then humanity will have unified morality. Therefore, if goodness is an objective entity and humanity does not have unified morality, then people cannot know about goodness.
$\neg(\alpha \rightarrow \beta) \rightarrow (\alpha \wedge \neg\beta)$	"If a person is immortal, then he will not die." B says: "But if this statement is false, then it means that there is someone who is immortal but has died."
$(\alpha \wedge \neg\beta) \rightarrow \neg(\alpha \rightarrow \beta)$	Xiao Ming works hard on homework but his grades are not excellent. Therefore, it is not the case that if he works hard on homework, then his grades will be excellent.

Table 1: Valid inferences and corresponding natural language examples.

from p to $\neg\neg$ p (see Figure 3 (Appendix E)). A parallel trend arises when we compare responses to two negation-related principles (Property $\neg(\phi \wedge \psi) \vdash \neg\phi \vee \neg\psi$) vs. Property $\neg\phi \vee \neg\psi \vdash \neg(\phi \wedge \psi)$; see Figure 4 (Appendix E)). LLMs often reject inferences that humans generally find no intuitive. For instance: from "There is no evidence that it will both rain and be windy tomorrow" to "There is no evidence that it will rain or no evidence that it will be windy tomorrow." These patterns suggest that LLMs tend not to overgeneralize Boolean negation to practical reasoning contexts. A plausible explanation is that negation is far more frequent in natural language corpora than other non-classical logical connectives. Thus, unlike with conditionals and modals(14) LLMs may not overgeneralize negation-based inference from limited patterns in training data.

Nevertheless, we also identified counterexamples in which LLM judgments diverge markedly from human judgments. For example, consider the common conversational negation: "I am not not respecting the teacher. Since you are not not respecting the teacher, you are respecting the teacher. Is the inference correct?" (We text in Chinese and it is grammatically correct). Most models answer no, despite the fact that this is a clear instance of classical double negation elimination. Even more striking, in linguistically richer examples such as the Chinese poem. For example "It is impossible that 'iron horses and frozen rivers invade my dreams'; therefore, either it is not 'iron horses and frozen rivers', or it does not invade my dreams. Is the inference valid?" —most LLMs continue to respond no. Taken together, these results reveal that, although LLMs sometimes align with human intuitions in resisting certain classically valid but pragmatically infelicitous inferences, they frequently fail to recognize genuinely valid classical negation inferences.

Our evaluation reveals notable patterns in LLM perfor-

mance concerning foundational negation principles. Specifically, LLMs exhibit a relatively high error rate on inference patterns derived from the Law of Excluded Middle, whereas their behavior on the principle of Explosion shows substantial variability: some models make errors at a strikingly high frequency, while others perform near-perfectly (see Figure 3 Appendix E). These observations further support our hypothesis that, rather than adhering to classical negation, LLMs more readily align with negation behavior characteristic of certain non-classical logics. To examine this hypothesis systematically, we aggregated results across logical systems and compared performance on inference patterns associated with classical logic against those associated with several non-classical frameworks. As shown in Table 6 and Figure 6 (Appendix F), the models consistently demonstrate stronger alignment with the negation principles of minimal logic, intuitionistic logic, and subminimal logic. This trend suggests that LLMs internalize a weaker and more context-sensitive interpretation of negation, one that departs from classical logic in a principled manner.

Robustness Evaluation

In pure logical reasoning, premises are expected to satisfy foundational structural properties of logical consequence—associativity, commutativity, contraction, and weakening:

$$(\varphi \wedge (\psi \wedge \chi)) \equiv ((\varphi \wedge \psi) \wedge \chi) \quad \text{(Associativity)}$$

$$(\varphi \wedge \psi) \equiv (\psi \wedge \varphi) \quad \text{(Commutativity)}$$

$$(\varphi \wedge \varphi) \vdash \varphi \quad \text{(Contraction)}$$

$$\varphi \vdash (\varphi \wedge \psi) \quad \text{(Weakening)}$$

Furthermore, predicate substitution should preserve logical validity: purely logical inference must remain invariant to changes in surface linguistic content. To systematically evaluate whether LLMs uphold these invariance and structural principles—and to reduce interference from natural-language context—we introduced four robustness interventions:

- **Concept Substitution.** Substitution of key predicates with irrelevant or fabricated concepts.
- **Redundancy and Negation.** Injection of logically redundant premises or paraphrastic negation.
- **Division and Recomposition.** Splitting, reordering, or recombining premises while preserving logical form.
- **Irrelevant Interference.** Addition of semantically loaded but logically irrelevant statements.

The following illustrate the types of modified stimuli used:

I. Concept Substitution

> "Since you said you are not disrespecting Jimmy, you must be respecting Jimmy." Is this inference correct? (**YES/NO**)

II. Redundancy and Negation

> "One plus one is either equal to two or not equal to two. Therefore, it is not the case that quantum wavefunction collapse is neither an objective nor a subjective process." Is this inference correct? (**YES/NO**)

III. Division and Recomposition

> "One cannot accept both 'If the moon rises in the east, then the sun rises in the east' and 'If the sun rises in the east, then the moon rises in the east'. Given also that if the moon rises in the west, the sun must rise in the west, does it follow that when the moon rises in the east, the sun rises in the west?" (**YES/NO**)

IV. Irrelevant Interference

> "Sugar-free drinks taste bad. If knowledge is infinite, then humans can know everything. But if this statement is false, then knowledge is infinite although humans cannot know everything." Is this reasoning acceptable? (**YES/NO**)

As reported in Table 7, 8 and Figure 5 (Appendix G), robustness interventions had a substantial influence on performance. Even the strongest models, such as Gemini 2.5 Pro, experienced accuracy reductions of nearly 50%, while DeepSeek V3.1 dropped to approximately 33% accuracy. In our robustness evaluation, Doubao exhibited the strongest performance among all tested models. Although its accuracy decreased by approximately 20% under perturbation, it still maintained an accuracy exceeding 60%, outperforming every other model in this setting. These outcomes highlight a critical limitation of current LLMs:

> *LLM reasoning is highly susceptible to superficial linguistic perturbations, indicating that they lack the stability characteristic of human-like logical competence.*

Conclusion and Limitations

In this study, we developed a benchmark grounded in a systematic framework of negation-based reasoning and evaluated the logical performance of fourteen large language models (LLMs). Our findings reveal several noteworthy patterns: (i) none of the evaluated models achieved an accuracy exceeding 90% on our tasks; (ii) although chain-of-thought prompting improves performance, the gains remain limited; (iii) LLMs appear to align more closely with non-classical treatments of negation than with classical logical principles; and (iv) the models exhibit substantial instability when subjected to perturbations in negation-related inference. Naturally, our work has several limitations. The test items, expressed in natural language, may inevitably introduce linguistic confounds; the size of the question set remains relatively modest; and our focus was restricted to negation reasoning, without isolating the potential influence of other logical constructs. Nevertheless, the present study provides a novel perspective for comparing inferential tendencies in LLMs and humans, particularly regarding patterns of negation. Future research will aim to broaden both the scope and

scale of the evaluation, thereby enabling a more comprehensive understanding of LLM logical capacities. Moreover, we have not yet addressed the question of how to enhance reasoning performance in LLMs—an important direction that we intend to pursue in subsequent work.

References

[1] A. Białynicki-Birula and H. Rasiowa, "On the representation of quasi-Boolean algebras," *Indagationes Mathematicae*, vol. 19, no. 4, pp. 403–410, 1957.

[2] S. Bowman, G. Angeli, C. Potts, and C. D. Manning, *A large annotated corpus for learning natural language inference*, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.

[3] F. Cheng, H. Li, F. Liu, R. van Rooij, K. Zhang, and Z. Lin, "Empowering LLMs with logical reasoning: A comprehensive survey," 2025 (to appear).

[4] K. Došen, "On negation," *Notre Dame Journal of Formal Logic*, vol. 35, no. 1, pp. 115–129, 1994.

[5] M. Dummett, *Elements of Intuitionism*. Oxford, U.K.: Oxford Univ. Press, 1977.

[6] J. M. Dunn, "Positive modal logic," *Studia Logica*, vol. 55, no. 2, pp. 301–317, 1995.

[7] J. M. Dunn, "Orthologic and quantum logic," in *Handbook of Philosophical Logic*, vol. 7, D. Gabbay and F. Guenthner, Eds., Dordrecht: Kluwer, 1995, pp. 3–107.

[8] J. St. B. T. Evans and D. E. Over, *If: Suppositions, Probability, and Reasoning*. Oxford University Press, 2004.

[9] P. Hájek, *Metamathematics of Fuzzy Logic*, Dordrecht: Kluwer, 1998.

[10] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, W. Zhou, J. Coady, D. Peng, Y. Qiao, L. Benson, L. Sun, A. Wardle-Solano, H. Szabo, E. Zubova, M. Burtell, J. Fan, Y. Liu, B. Wong, M. Sailor, A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, A. R. Fabbri, W. Kryściński, S. Yavuz, Y. Zhou, C. Xiong, R. Ying, A. Cohan, and D. Radev, "FOLIO: Natural Language Reasoning with First-Order Logic," In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pp. 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] S. Han, A. Yu, R. Shen, Z. Qi, M. Riddell, W. Zhou, Y. Qiao, Y. Zhao, S. Yavuz, Y. Liu, S. Joty, Y. Zhou, C. Xiong, D. Radev, R. Ying, and A. Cohan, "P-FOLIO: Evaluating and Improving Logical Reasoning with Abundant Human-Written Reasoning Chains," In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16553–16565, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] A. P. Hazen, "A logic of the weak negation," *Notre Dame Journal of Formal Logic*, vol. 31, no. 3, pp. 346–352, 1990.

[13] A. Heyting, *Intuitionism: An Introduction*, 2nd ed. Amsterdam, The Netherlands: North-Holland, 1966.

[14] W. H. Holliday, M. Mandelkern, and C. E. Zhang, "Conditional and Modal Reasoning in Large Language Models," In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3800–3821, Miami, Florida, USA, November 2024. Association for Computational Linguistics. .

[15] I. Johansson, "Der Minimalkalkül, ein reduzierter intuitionistischer Formalismus," *Compositio Mathematica*, vol. 4, pp. 119–136, 1937.

[16] J. Kalman, "Lattices with involution," *Transactions of the American Mathematical Society*, vol. 87, no. 2, pp. 485–491, 1958.

[17] S. C. Kleene, *Introduction to Metamathematics*, New York: Van Nostrand, 1952.

[18] T. Kojima, A. Gu, J. Reif, and Y. Shin, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[19] A. N. Kolmogorov, "On the principle of excluded middle," in *Mathematics and Logic: Foundations of Mathematics*, vol. 1, 1925, pp. 414–437.

[20] A. Leidinger, R. Van Rooij, et al., *Are LLMs Classical or Non-monotonic Reasoners? Lessons from Generics*, ACL, 2024.

[21] J. Łukasiewicz, "On three-valued logic," *Ruch Filozoficzny*, vol. 5, pp. 169–171, 1920.

[22] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, *Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning*, ICLR, 2024, arXiv:2310.01061.

[23] G. C. Moisil, "Recherches sur les logiques non chrysippiennes," *Annales Scientifiques de l'Université de Jassy*, vol. 28, pp. 1–30, 1942.

[24] T. Morishita et al., *Enhancing Reasoning Capabilities of LLMs via Principled Synthetic Logic Corpus*, NeurIPS, 2024.

[25] D. Nelson, "Constructible falsity," *The Journal of Symbolic Logic*, vol. 14, no. 1, pp. 16–26, 1949.

[26] Y. Nie, A. Williams, E. Gardner, N. Nangia, and S. Bowman, *Adversarial NLI: A new benchmark for natural language understanding*, NAACL, 2019.

[27] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[28] P. Portner, *Modality*. Oxford University Press, 2009.

[29] D. Prawitz, *Natural Deduction: A Proof-Theoretical Study*. Stockholm, Sweden: Almqvist & Wiksell, 1965.

[30] G. Priest, *In Contradiction*, 2nd ed., Oxford: Oxford University Press, 2006.

[31] A. Saparov and X. He, *Title of Saparov and He 2023 paper*, Conference/Journal Name, 2023.

[32] O. Tafjord, B. Dalvi Mishra, and P. Clark, "ProofWriter: Generating implications, proofs, and abductive statements over natural language," *Findings*

of the Association for Computational Linguistics (ACL/IJCNLP), pages 3621–3634, 2021.

[33] A. Tarski, "On the concept of logical consequence," *Logic, Semantics, Metamathematics*, pp. 409–420, 1936.

[34] J.Tian, Y.Li, W.Chen, L.Xiao, H.He, and Y.Jin, "Diagnosing the First-Order Logical Reasoning Ability Through LogicNLI," In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.3738–3747, 2021.

[35] A. S. Troelstra and D. van Dalen, *Constructivism in Mathematics: An Introduction*, vol. 1. Amsterdam, The Netherlands: North-Holland, 1988.

[36] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Chi, H. Le, X. Zhou, M. Chung, P. Rao, C. Song, S. Zhang, B. Wang, B. Zhou, J. Tang, S. Gu, P. Li, T. Zhou, D. Wang, H. Li, A. Amodei, and S. Liang, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.

[37] A. Williams, N. Nangia, and S. Bowman, *A broad-coverage challenge corpus for sentence understanding through inference*, NAACL, 2018.

Appendix A

In this section, we list all the LLMs used in our experiments: the DeepSeek-V3.2-Exp, DeepSeek-V3.1, and DeepSeek-V3 models (DeepSeek); the Moonshot-Kimi-K2-Instruct model (Moonshot AI); the Doubao-Seed-1.6 and Doubao-1.5-UI-TARS models (ByteDance — Doubao Team); the Qwen-Plus-Latest and Qwen3-Next-80B-A3B-Instruct models (Alibaba Cloud — Qwen Team); the ChatGLM-4.6 and ChatGLM-4.5 models (Zhipu AI); the Ernie Bot 4.5 Turbo model (Baidu — ERNIE Team); the Gemini 2.5 Flash and Gemini 2.5 Pro models (Google DeepMind); and the GPT-5 mini model (OpenAI).

Appendix B

We tested each of the following properties with a series of examples, and only a selection of them is shown here. The results are shown in Figures 1.

$\neg\neg\alpha \rightarrow \alpha$

Example 1 *Does "It is impossible to prove that we cannot find a proof of this problem" mean that we can find a proof of this problem?*
If yes, answer "YES". If no, answer "NO".

$\alpha \rightarrow \neg\neg\alpha$

Example 2 *If this system is indeed inconsistent, does that mean we cannot prove the conclusion "we cannot assert that it is inconsistent"?*
If yes, answer "YES". If no, answer "NO".
Note: You should answer this question and only this question with "YES" or "NO".

$\alpha \wedge \neg\alpha \rightarrow \beta$ **(Principle of Explosion)**

Example 3 *The philosopher said: "If a person is both kind and cruel at the same time, then he possesses ultimate moral perfection." The audience was shocked: "That's a madman's theory."*
Is the philosopher's statement correct? If your judgment is yes, answer "YES"; otherwise, answer "NO".
Note: You should answer this question and only this question with "YES" or "NO".

$\alpha \wedge \neg\alpha \rightarrow \bot$

Example 4 *It was the best of times, it was the worst of times. So false is true. Is this conclusion valid? Answer "YES" or "NO". Note, answer this question and only answer "YES" or "NO".*

$\beta \rightarrow (\alpha \vee \neg\alpha)$

Example 5 *Philosopher: "Every action is either good or evil."*
Audience responded: "There are also neutral or unintended actions."
Is the philosopher's statement correct? If your judgment is yes, answer "YES"; otherwise, answer "NO".
Note: You should answer this question and only this question with "YES" or "NO".

$\top \rightarrow (\alpha \vee \neg\alpha)$

Example 6 *Because it is either the weekend or not the weekend, therefore "either the quantum state collapses or the quantum state remains in superposition". Is the above derivation correct? If you determine that this derivation is valid, choose "YES"; if it is not valid, choose "NO". Note, answer this question and only answer "YES" or "NO".*

If $\alpha \rightarrow \beta$**, then** $\neg\beta \rightarrow \neg\alpha$

Example 7 *The teacher told the student: "If you complete all the homework, then you will pass the exam." The student did not pass the exam, so he did not complete all the homework. Is this derivation correct? If correct, choose "YES"; if incorrect, choose "NO". Note, answer this question and only answer "YES" or "NO".*

If $\alpha \rightarrow \neg\beta$ **, then** $\beta \rightarrow \neg\alpha$

Example 8 *If we are an honest person, then we cannot meet people's expectations. Therefore, if we have met people's expectations, we have actually not been an honest person. Is this reasoning correct? If yes, answer "YES"; if no, answer "NO". Note, answer this question and only answer "YES" or "NO".*

$\neg(\alpha \wedge \beta) \rightarrow (\neg\alpha \vee \neg\beta)$

Example 9 *The parent said to the child: "You cannot both stay up late every day and maintain a failing grade." The child thought: "What Mom means is that I can fail, but I shouldn't stay up late every day." Is the child's understanding correct? If your judgment is "correct", choose "YES"; otherwise, choose "NO". Note, answer this question and only answer "YES" or "NO".*

$(\neg\alpha \vee \neg\beta) \rightarrow \neg(\alpha \wedge \beta)$

Example 10 *The officer said: "With insufficient evidence, we cannot conclude that Zhang San is not the murderer or that Li Si is not the murderer. There are too many unresolved clues." So, is the officer concluding that "not both are murderers"? If your judgment is "yes", choose "YES"; otherwise, choose "NO". Note, answer this question and only answer "YES" or "NO".*

$(\neg\alpha \wedge \neg\beta) \rightarrow \neg(\alpha \vee \beta)$

Example 11 *The island is neither in the Southern Hemisphere nor in the Northern Hemisphere. Therefore, it is not the case that the island is in the Southern Hemisphere or the Northern Hemisphere. Is the above reasoning correct? If yes, answer "YES"; if no, answer "NO". Note, answer this question and only answer "YES" or "NO".*

$\neg(\alpha \vee \beta) \rightarrow (\neg\alpha \wedge \neg\beta)$

Example 12 *The philosopher said: "If there is no beauty or truth in the world, then there is no beauty and no truth in the world." Question: Is this statement true? Note, answer this question and only answer "YES" or "NO".*

If $\alpha \wedge \beta \rightarrow r$**, then** $\alpha \wedge \neg r \rightarrow \neg\beta$

Example 13 *Xiao Ming's mother said to him: "If the store has it in stock and you behave well, then I will buy you badminton." It is known that the store did have it in stock, but Xiao Ming's mother did not have enough money that day and did not buy the badminton. From this, it is inferred that Xiao Ming did not behave well that day. Is this reasoning correct? If you determine that this reasoning is correct, choose "YES"; if you determine that this reasoning is incorrect, choose "NO". Note, answer this question and only answer "YES" or "NO".*

$\neg(\alpha \rightarrow \beta) \rightarrow (\alpha \wedge \neg\beta)$

Example 14 *Xiao Li said, "If Zhang San comes to the party, then Li Si will also come." Xiao Zhao said: "That's not right, that's not the case. I've long known that Li Si dislikes Wang Wu, and Wang Wu, as the organizer, will definitely attend the party. Therefore, whether Zhang San comes or not, Li Si will not come."*

So, does Xiao Zhao believe that Zhang San will come, but Li Si will not come? If your judgment is "yes", choose "YES"; otherwise, choose "NO". Note, answer this question and only answer "YES" or "NO".

$(\alpha \wedge \neg\beta) \rightarrow \neg(\alpha \rightarrow \beta)$

Example 15 *He reviewed on time and did not get a high score. Therefore, it is not the case that if he reviews on time, he will get a high score. If you determine this derivation is valid, choose "YES"; if not, choose "NO".*

Appendix C

In response to the following question, think according to the following chain before you answer:

First, precisely identify and define the core problem. Clarify its essence and boundaries, and define all key terms to avoid misinterpretation.

Second, decompose the problem systematically and analyze it from multiple perspectives. Break it into subproblems or dimensions; examine it in terms of pros and cons, as well as internal and external causes; and gather relevant principles and factual evidence.

Third, review the relevant logical principles and reason accordingly. Ensure each step follows from the previous one; make explicit and evaluate each step's logical and causal links; construct a complete chain of inference; and draw conclusions cautiously, stating their premises.

Finally, based on the above analysis, derive the final answer or solution, and clearly state the conditions under which the conclusion applies and its possible limitations.

Appendix D

The results are shown in Tables 2, 3 4, 5, and Figure 2.

Appendix E

The results are shown in Figures 3 and 4.

Appendix F

The results are shown in Table 6 and Figure 6.

Appendix G

The results are shown in Tables 7, 8, and Figure 5.

tem	Version	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average Accuracy
0	DeepSeek-V3.2-Exp	13%	81%	17%	3%	30%	83%	83%	100%	73%	100%	100%	100%	83%	71%	100%	69%
0	DeepSeek-V3.1	81%	83%	90%	91%	70%	79%	86%	92%	81%	89%	97%	87%	90%	88%	91%	86%
0	DeepSeek-V3	33%	100%	0%	29%	17%	0%	77%	83%	90%	100%	77%	100%	50%	100%	100%	64%
0	Moonshot-Kimi-K2-Instruct	53%	100%	0%	0%	3%	0%	77%	100%	87%	100%	100%	100%	60%	71%	100%	63%
0	Doubao-Seed-1.6	43%	72%	60%	100%	43%	83%	87%	97%	90%	97%	100%	97%	83%	100%	100%	83%
0	Doubao-1.5-UI-TARS	50%	84%	17%	32%	40%	83%	90%	100%	80%	97%	100%	84%	87%	97%	100%	76%
0	Qwen-Plus-Latest	77%	97%	40%	32%	33%	83%	43%	97%	90%	94%	77%	87%	47%	81%	100%	72%
0	Qwen3-Next-80B-A3B-Instruct	50%	69%	0%	13%	23%	67%	50%	87%	57%	84%	32%	52%	27%	74%	87%	51%
0	ChatGLM-4.6	7%	100%	90%	97%	50%	87%	63%	50%	73%	65%	81%	90%	67%	100%	100%	75%
0	ChatGLM-4.5	7%	100%	83%	100%	30%	100%	87%	70%	67%	97%	100%	100%	83%	100%	100%	82%
0	Ernie Bot4.5 Turbo	60%	84%	0%	0%	0%	0%	7%	83%	73%	100%	97%	71%	83%	19%	50%	49%
0	Gemini2.5flash	13%	91%	90%	97%	37%	17%	80%	97%	7%	84%	97%	74%	83%	87%	93%	70%
0	Gemini2.5pro	23%	78%	97%	97%	60%	93%	90%	100%	93%	65%	94%	81%	87%	87%	93%	82%
0	GPT-5 mini	47%	100%	3%	26%	57%	17%	27%	97%	0%	87%	3%	97%	73%	97%	0%	49%

Table 2: Accuracy0

tem	Version	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average Accuracy
1	DeepSeek-V3.2-Exp	20%	69%	50%	26%	37%	17%	97%	83%	97%	97%	81%	100%	50%	100%	100%	68%
1	DeepSeek-V3.1	82%	83%	88%	92%	72%	80%	86%	93%	80%	88%	95%	89%	91%	87%	92%	87%
1	DeepSeek-V3	33%	100%	0%	32%	17%	90%	57%	77%	90%	97%	90%	97%	67%	97%	100%	70%
1	Moonshot-Kimi-K2-Instruct	77%	100%	0%	0%	3%	63%	67%	83%	67%	97%	97%	100%	67%	100%	33%	64%
1	Doubao-Seed-1.6	53%	72%	67%	84%	43%	70%	80%	83%	60%	71%	84%	84%	43%	81%	97%	71%
1	Doubao-1.5-UI-TARS	50%	81%	20%	39%	20%	83%	90%	100%	80%	97%	97%	90%	87%	97%	100%	75%
1	Qwen-Plus-Latest	77%	100%	0%	100%	0%	100%	90%	100%	67%	87%	45%	100%	60%	84%	100%	74%
1	Qwen3-Next-80B-A3B-Instruct	50%	69%	0%	13%	23%	67%	57%	87%	57%	84%	32%	52%	27%	74%	87%	52%
1	ChatGLM-4.6	13%	100%	90%	71%	47%	40%	33%	43%	53%	65%	29%	0%	70%	0%	100%	50%
1	ChatGLM-4.5	17%	100%	17%	39%	0%	0%	67%	83%	87%	35%	84%	52%	87%	65%	33%	51%
1	Ernie Bot4.5 Turbo	87%	81%	0%	0%	0%	0%	13%	83%	80%	100%	97%	100%	0%	26%	50%	48%
1	Gemini2.5flash	27%	81%	90%	97%	20%	83%	73%	97%	63%	97%	97%	68%	83%	84%	93%	77%
1	Gemini2.5pro	30%	84%	83%	87%	30%	60%	87%	63%	70%	68%	29%	55%	97%	65%	60%	65%
1	GPT-5 mini	57%	78%	17%	35%	60%	20%	40%	70%	3%	42%	10%	74%	53%	61%	10%	42%

Table 3: Accuracy1

Version	tem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average
DeepSeek-V3.2-Exp	0	10%	88%	21%	7%	32%	93%	93%	100%	73%	100%	100%	100%	92%	75%	100%	72%
DeepSeek-V3.1	0	76%	81%	87%	88%	86%	87%	94%	100%	83%	100%	92%	100%	79%	92%	97%	89%
DeepSeek-V3	0	37%	100%	0%	23%	17%	17%	80%	100%	93%	97%	94%	97%	77%	97%	100%	68%
Moonshot-Kimi-K2-Instruct	0	62%	100%	8%	9%	5%	12%	84%	100%	90%	98%	100%	100%	71%	81%	100%	68%
Doubao-Seed-1.6	0	44%	80%	74%	100%	46%	94%	90%	100%	100%	98%	100%	100%	84%	100%	100%	87%
Doubao-1.5-UI-TARS	0	55%	84%	19%	47%	47%	95%	97%	100%	82%	100%	100%	91%	100%	100%	100%	81%
Qwen-Plus-Latest	0	89%	100%	50%	41%	38%	88%	51%	100%	90%	93%	81%	98%	48%	94%	100%	77%
Qwen3-Next-80B-A3B-Instruct	0	57%	64%	4%	14%	26%	71%	46%	82%	64%	87%	34%	49%	30%	79%	95%	53%
ChatGLM-4.6	0	11%	100%	100%	100%	62%	86%	71%	58%	73%	77%	92%	95%	81%	100%	99%	80%
ChatGLM-4.5	0	20%	100%	86%	100%	29%	100%	93%	80%	72%	100%	99%	100%	90%	99%	100%	85%
Ernie Bot4.5 Turbo	0	68%	91%	9%	4%	6%	6%	9%	92%	73%	96%	100%	70%	86%	25%	58%	53%
Gemini2.5flash	0	26%	99%	98%	100%	36%	29%	83%	100%	7%	95%	96%	79%	88%	99%	100%	76%
Gemini2.5pro	0	32%	87%	100%	100%	75%	100%	94%	100%	100%	66%	100%	87%	95%	100%	100%	89%
GPT-5 mini	0	46%	100%	7%	34%	62%	23%	36%	100%	7%	96%	16%	100%	87%	100%	13%	55%

Table 4: Chain of Thought T0

Version	tem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average
DeepSeek-V3.2-Exp	1	33%	69%	60%	24%	42%	22%	100%	91%	100%	98%	79%	99%	54%	100%	99%	71%
DeepSeek-V3.1	1	77%	83%	86%	90%	85%	85%	93%	98%	81%	99%	90%	97%	80%	90%	95%	89%
DeepSeek-V3	1	67%	100%	0%	29%	17%	67%	97%	100%	93%	97%	97%	97%	77%	97%	100%	76%
Moonshot-Kimi-K2-Instruct	1	85%	98%	13%	11%	7%	67%	66%	92%	78%	100%	100%	100%	72%	100%	47%	69%
Doubao-Seed-1.6	1	53%	75%	67%	87%	45%	84%	93%	90%	67%	70%	87%	84%	58%	92%	100%	77%
Doubao-1.5-UI-TARS	1	59%	89%	31%	45%	26%	82%	100%	99%	82%	100%	97%	93%	93%	100%	100%	80%
Qwen-Plus-Latest	1	83%	100%	11%	99%	12%	100%	89%	100%	78%	86%	56%	100%	61%	90%	100%	78%
Qwen3-Next-80B-A3B-Instruct	1	55%	76%	15%	16%	31%	68%	64%	91%	61%	93%	45%	52%	34%	79%	94%	58%
ChatGLM-4.6	1	19%	100%	88%	78%	49%	46%	36%	52%	59%	68%	43%	13%	81%	10%	100%	56%
ChatGLM-4.5	1	25%	100%	25%	40%	16%	14%	81%	97%	91%	51%	92%	59%	89%	70%	47%	60%
Ernie Bot4.5 Turbo	1	94%	89%	3%	8%	3%	1%	13%	89%	76%	96%	100%	98%	2%	23%	46%	50%
Gemini2.5flash	1	30%	94%	100%	100%	33%	94%	78%	100%	81%	100%	100%	74%	96%	88%	100%	85%
Gemini2.5pro	1	33%	94%	93%	87%	33%	63%	95%	61%	80%	68%	36%	61%	100%	74%	57%	69%
GPT-5 mini	1	58%	81%	19%	38%	72%	28%	39%	76%	6%	46%	13%	85%	61%	72%	12%	47%

Table 5: Chain of Thought T1

Model	Intuitionistic Logic	Minimal Logic	Orthon Negation	Subminimal Logic	De Morgan
DeepSeek-V3.2-Exp	71%	79%	64%	67%	71%
DeepSeek-V3.1	90%	89%	89%	91%	88%
DeepSeek-V3	65%	74%	61%	58%	71%
Moonshot-Kimi-K2-Instruct	67%	77%	66%	59%	75%
Doubao-Seed-1.6	87%	91%	82%	93%	85%
Doubao-1.5-UI-TARS	74%	82%	72%	77%	79%
Qwen-Plus-Latest	65%	69%	66%	50%	72%
Qwen3-Next-80B-A3B-Instruct	41%	47%	42%	31%	49%
ChatGLM-4.6	80%	78%	72%	77%	70%
ChatGLM-4.5	90%	91%	81%	93%	79%
Ernie Bot4.5 Turbo	53%	61%	54%	47%	62%
Gemini2.5flash	89%	88%	80%	89%	71%
Gemini2.5pro	91%	90%	83%	92%	83%
GPT-5 mini	53%	60%	53%	32%	52%

Table 6: Accuracy of models under various non-classical logics.

	Version	Concept Replacement (53% to 40%)								Redundancy and Negation (56% to 27%)						Splitting and Recombination (77% to 56%)								Irrelevant Interference (71% to 52%)							
		1		2		3		4		5		6		7		8		9		10		11		12		13		14		15	
		pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post
DeepSeek	DeepSeek-V3.2-Exp DeepSeek-V3.1 DeepSeek-V3	47%	12%	92%	55%	43%	0%	48%	15%	50%	0%	65%	0%	83%	7%	91%	27%	89%	47%	99%	47%	91%	45%	99%	46%	75%	34%	95%	47%	100%	47%
Kimi	Moonshot-Kimi-K2-Instruct	65%	56%	100%	98%	0%	0%	0%	6%	3%	0%	32%	12%	72%	6%	92%	84%	77%	56%	98%	58%	98%	54%	100%	100%	63%	72%	85%	17%	67%	58%
Doubao	Doubao-Seed-1.6 Doubao-1.5-UI-TARS	49%	57%	77%	69%	41%	67%	64%	84%	37%	40%	80%	67%	87%	10%	95%	7%	78%	57%	90%	84%	95%	84%	89%	81%	75%	53%	94%	84%	99%	93%
Qwen	Qwen-Plus-Latest Qwen3-Next-80B-A3B-Instruct	63%	21%	84%	35%	10%	6%	40%	40%	20%	1%	79%	37%	60%	7%	93%	47%	68%	29%	87%	29%	47%	32%	73%	65%	40%	39%	78%	51%	93%	91%
ChatGLM	ChatGLM-4.6 ChatGLM-4.5	11%	17%	100%	100%	70%	17%	77%	29%	32%	0%	57%	0%	63%	67%	62%	83%	70%	67%	65%	84%	73%	0%	60%	0%	77%	83%	66%	65%	83%	100%
Ernie Bot	Ernie Bot4.5 Turbo	73%	87%	83%	66%	0%	0%	0%	0%	0%	0%	0%	0%	10%	13%	83%	87%	77%	80%	100%	97%	97%	100%	85%	97%	42%	0%	23%	0%	50%	50%
Gemini	Gemini2.5flash Gemini2.5pro	23%	4%	84%	82%	90%	36%	94%	86%	37%	25%	63%	42%	83%	8%	89%	73%	58%	65%	78%	47%	79%	47%	69%	54%	88%	32%	81%	42%	85%	54%
ChatGPT	GPT-5 mini	52%	16%	89%	90%	10%	5%	31%	15%	58%	31%	18%	7%	33%	14%	83%	76%	2%	2%	65%	63%	6%	0%	85%	63%	63%	53%	79%	72%	5%	4%

Table 7: Robustness Testing

Model	Accuracy (%)		
($T = 0$)	Overall Accuracy	CoT Accuracy	Robustness Accuracy
DeepSeek-V3.2-Exp	69.25	72.22	29.20
DeepSeek-V3.1	86.33	89.48	
DeepSeek-V3	63.76	68.43	
Moonshot-Kimi-K2-Instruct	63.40	68.07	45.37
Doubao-Seed-1.6	83.47	87.26	63.23
Doubao-1.5-UI-TARS	76.05	81.12	
Qwen-Plus-Latest	71.86	77.38	36.90
Qwen3-Next-80B-A3B-Instruct	51.35	53.44	
ChatGLM-4.6	74.59	80.46	49.58
ChatGLM-4.5	81.56	84.56	
Ernie Bot4.5 Turbo	48.54	52.75	43.31
Gemini2.5flash	69.73	75.71	46.23
Gemini2.5pro	82.49	89.05	
GPT-5 mini	48.65	55.13	34.63

Table 8: Accuracy

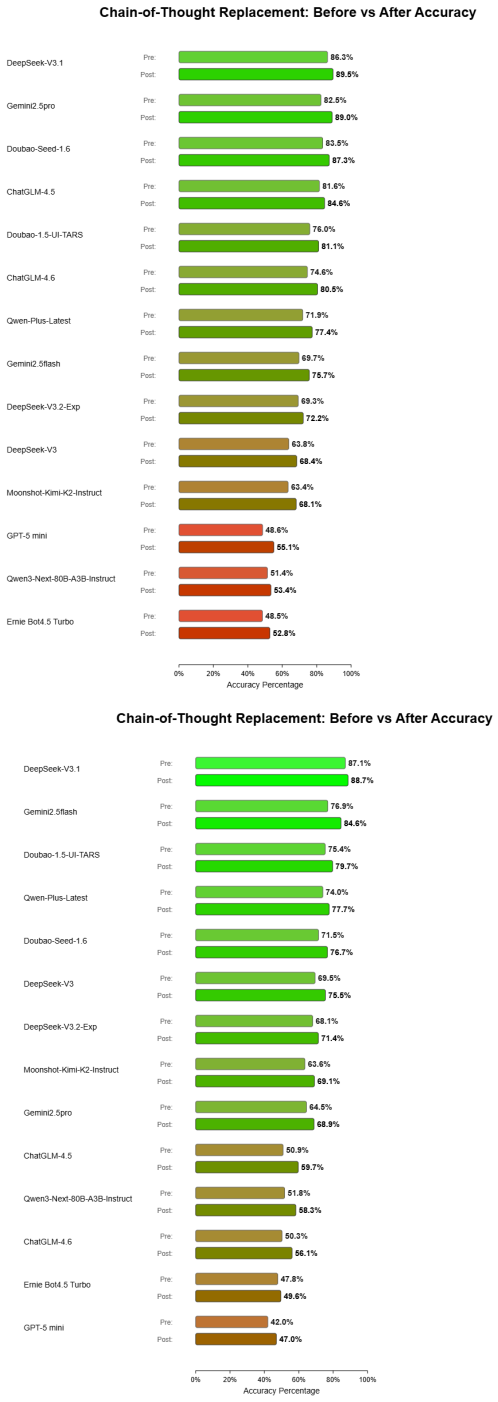
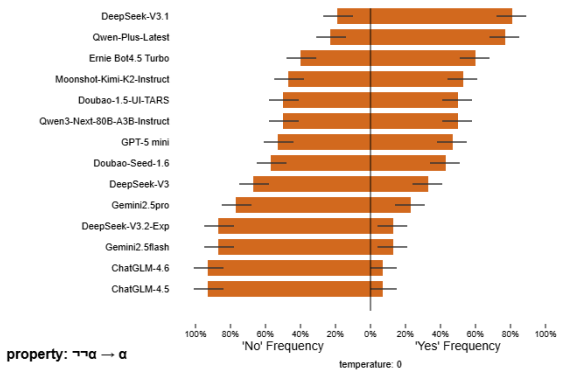
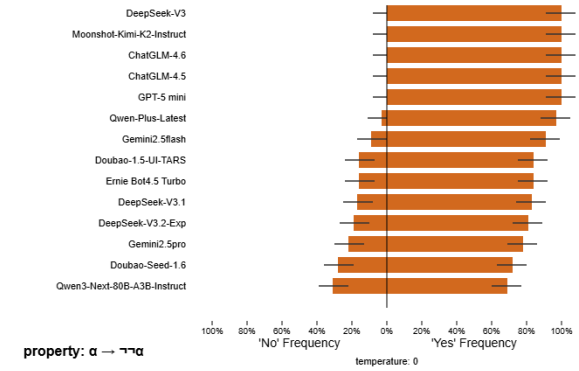
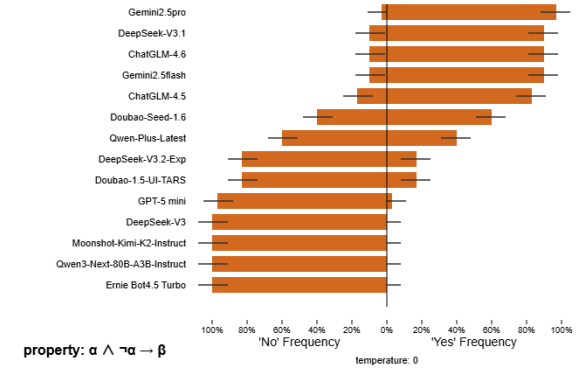


Figure 2: Chain-of-Thought Replacement-Before vs After Accuracy







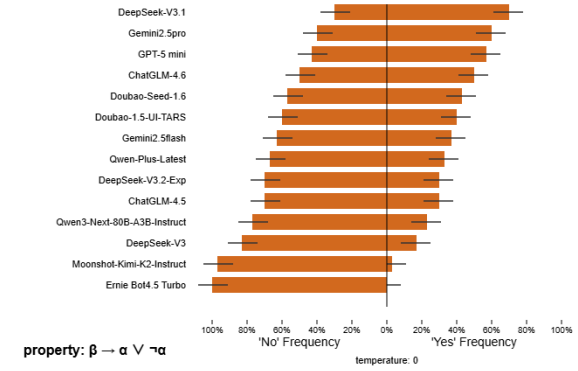


Figure 3: Properties

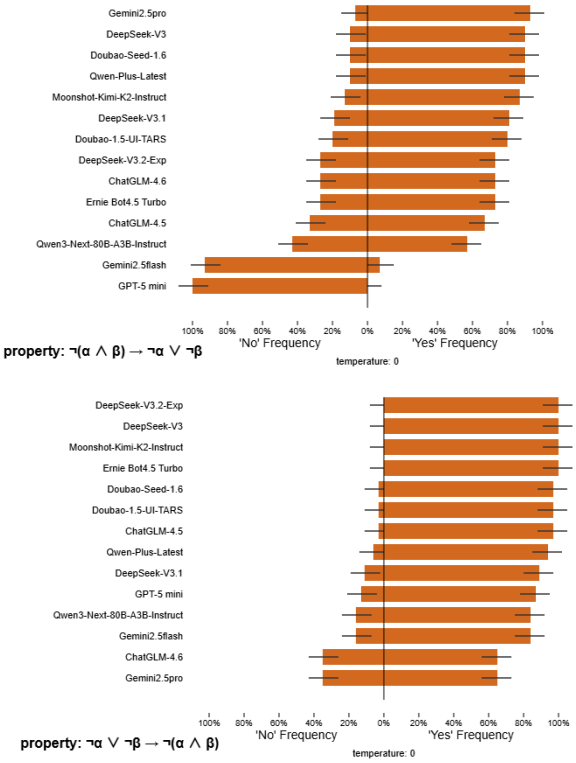


Figure 4: Properties

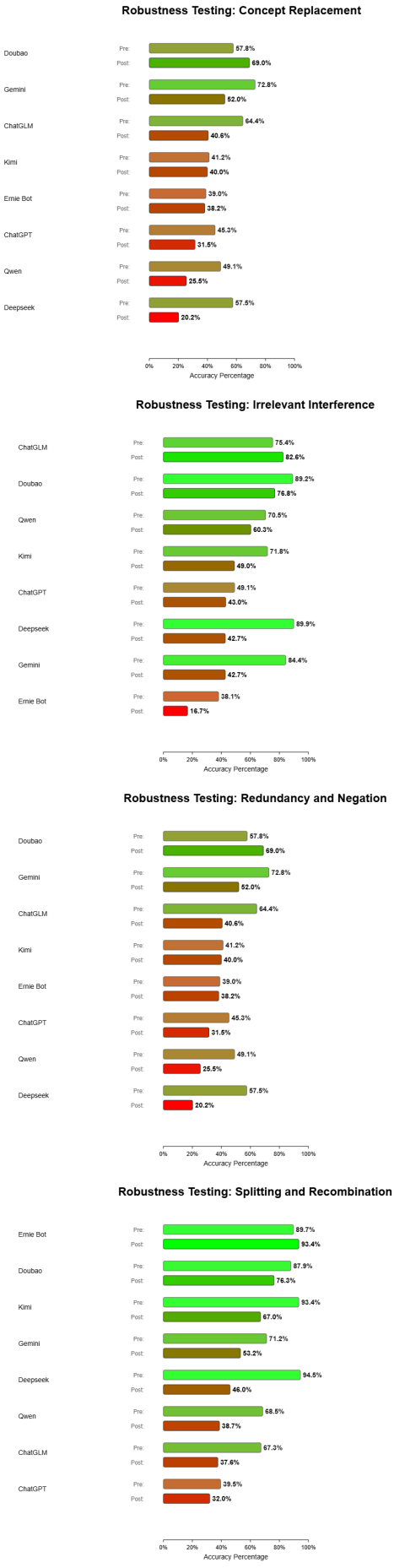


Figure 5: Robustness

Figure 6: Logic Comparison