MODALBENCH: EVALUATING MODAL AND DEONTIC LOGIC REASONING IN LARGE LANGUAGE MODELS

Mujtaba Hasan
New Delhi, India
mujtaba.hasan@live.com

ABSTRACT

Large language models excel at propositional and first-order logic, yet no benchmark tests whether they can reason about *necessity*, *possibility*, *obligation*, or *permission*—the domain of modal and deontic logic. We introduce **ModalBench**, the first benchmark grounded in Kripke semantics for evaluating modal reasoning in LLMs. ModalBench consists of 1,500 problems spanning five modal systems (K, T, S4, S5, and deontic D), three difficulty tiers, and two presentation tracks (formal symbolic and natural language), each with computationally verified ground truth. We evaluate three LLMs with three prompting strategies (27,000 total inferences) and find: (1) under chain-of-thought, models achieve 73–91% accuracy, with system K (unconstrained accessibility) as the hardest at 65–75%; (2) **LLMs reason about modality better in natural language than in formal notation**—two of three models score significantly higher on narrative presentations, with Qwen3-235B showing an 18.4 pp advantage for natural language ($p < 10^{-99}$); (3) **world-enumeration prompting acts as a scaffold**: it boosts weaker models by 9–11 pp but is unnecessary for the strongest model, which peaks at 95.9% with zero-shot prompting; (4) all models exhibit **implicit S5 bias**, over-applying the Euclidean axiom by up to $+56$ pp in system K where it is not valid. We release ModalBench, the evaluation code, and all results at https://github.com/mujtabahasan/modalbench.

1 INTRODUCTION

When a doctor says a patient "must have an infection," or a contract states that a party "may terminate upon notice," the reasoning involved is *modal*: it concerns not just what is true, but what is *necessarily* true, *possibly* true, *obligatory*, or *permitted*. Modal logic formalizes these notions with operators $\Box$ (necessity) and $\Diamond$ (possibility), evaluated over possible worlds connected by accessibility relations (Kripke, 1963; Chellas, 1980). Deontic logic extends this framework to normative reasoning with obligation (OB), permission (PE), and prohibition (FO) (McNamara & Van De Putte, 2021). These logics underpin applications in medical diagnosis, legal reasoning, AI safety, and planning under uncertainty.

Despite the importance of modal reasoning, existing LLM logic benchmarks—FOLIO (Han et al., 2024), LogiQA (Liu et al., 2020), ProofWriter (Tafjord et al., 2021), PrOntoQA (Saparov & He, 2023), LogicBench (Parmar et al., 2024)—evaluate only propositional and first-order logic. Theory-of-mind tasks implicitly probe epistemic modality but lack formal grounding, and legal benchmarks (Guha et al., 2024) involve deontic reasoning without isolating it. The survey by Cheng et al. (2025) explicitly identifies "extending to modal logic" as an open direction. We fill this gap.

Contributions. (1) ModalBench: the first benchmark with Kripke-semantic ground truth for five modal systems and three difficulty tiers. (2) A dual-track design (formal + natural language) revealing that most models reason *better* in natural language—challenging the assumption that symbolic presentation is inherently easier. (3) A prompting study showing world-enumeration acts as a scaffold for weaker models but not the strongest. (4) Per-axiom diagnostics uncovering systematic implicit S5 bias. (5) Five classical deontic paradoxes as discriminative test cases.

Table 1: Modal systems in ModalBench. Each constrains the accessibility relation R.

System	Constraint	Key Axiom	Intuition
K	None	$\Box(P \to Q) \to (\Box P \to \Box Q)$	Distribution
T	Reflexive	$\Box P \to P$	Necessity implies truth
S4	Refl. + Transitive	$\Box P \to \Box\Box P$	Positive introspection
S5	Equivalence	$\Diamond P \to \Box\Diamond P$	If possible, nec. possible
D	Serial	$\text{OB}(P) \to \text{PE}(P)$	Ought implies permitted

2 Background: Kripke Semantics

A **Kripke model** $\mathcal{M} = \langle W, R, V \rangle$ consists of possible worlds W, an accessibility relation $R \subseteq W \times W$, and a valuation V assigning truth values to propositions at each world. The modal operators are:

$$\mathcal{M}, w \models \Box\varphi \iff \forall w' \colon (w, w') \in R \implies \mathcal{M}, w' \models \varphi \tag{1}$$

$$\mathcal{M}, w \models \Diamond\varphi \iff \exists w' \colon (w, w') \in R \ \wedge \ \mathcal{M}, w' \models \varphi \tag{2}$$

That is, $\Box\varphi$ holds at w when φ is true at *every* accessible world, and $\Diamond\varphi$ when φ is true at *some* accessible world. When no worlds are accessible, $\Box\varphi$ is vacuously true and $\Diamond\varphi$ is false.

Different constraints on R yield different modal systems (Table 1). A key test of modal competence is *axiom awareness*: knowing that axiom T ($\Box P \to P$) holds in system T (reflexive frames) but not in system K. We define the **implicit S5 bias** as the tendency to apply axioms valid only in S5—where R is an equivalence relation—even in weaker systems.

In deontic logic, $\Box/\Diamond$ become OB/PE. Standard deontic logic produces well-known paradoxes: **Ross's Paradox** ($\text{OB}(p) \to \text{OB}(p \vee q)$—valid but counterintuitive), **Chisholm's Paradox** (inconsistent contrary-to-duty norms), and the **Gentle Murderer** ("if you murder, you ought to murder gently").

2.1 Running Example

Consider a Kripke model in system T (reflexive) with three worlds $W = \{w_0, w_1, w_2\}$:

Accessibility. $R = \{(w_0, w_0), (w_0, w_1), (w_1, w_1), (w_1, w_2), (w_2, w_2)\}$.

Valuations. w_0: $p = \top, q = \bot$ | w_1: $p = \bot, q = \top$ | w_2: $p = \top, q = \top$.

Tier 1: Is $\Box p$ true at w_0? Accessible worlds: $\{w_0, w_1\}$. Since p is false at w_1, $\Box p$ is **False**—p is not true at *all* accessible worlds.

Tier 1: Is $\Diamond(p \wedge q)$ true at w_0? Check $\{w_0, w_1\}$: $(p \wedge q)_{w_0} = \top \wedge \bot = \bot$; $(p \wedge q)_{w_1} = \bot \wedge \top = \bot$. No accessible world satisfies $p \wedge q$, so $\Diamond(p \wedge q)$ is **False**.

Tier 3: Is $\Box p \to p$ true at w_0? We showed $\Box p = \bot$. Since $\bot \to$ anything $= \top$, the formula is **True**—vacuously. This illustrates a critical subtlety: axiom T is always true in reflexive frames, but its truth often follows from a false antecedent, testing whether models understand the material conditional.

2.2 Deontic Paradoxes in Detail

Standard deontic logic (SDL) produces paradoxes that expose limitations of the formalism and serve as highly discriminative benchmark problems:

Ross's Paradox. $\text{OB}(p) \to \text{OB}(p \vee q)$: if you must mail a letter, SDL says you must "mail it or burn it." This follows from monotonicity of OB and tests whether models apply it despite the counterintuitive conclusion.

Chisholm's Paradox. Consider: (i) $\text{OB}(h)$, (ii) $h \to \text{OB}(t)$, (iii) $\neg h \to \text{OB}(\neg t)$, (iv) $\neg h$. Rules (i)–(iv) yield contradiction in SDL. This tests detection of inconsistency in normative rule sets—critical for legal reasoning.

Gentle Murderer. "If you murder, you ought to murder gently": a contrary-to-duty obligation that SDL cannot elegantly represent because $\text{OB}(\neg\text{murder})$ conflicts with $\text{OB}(\text{gently} \mid \text{murder})$. All models score below 79% on this (§5.6).

FO/PE Contradiction. $\text{FO}(p) \wedge \text{PE}(p)$—forbidden yet permitted—is always False in serial frames. Tests basic deontic consistency.

Good Samaritan. $\text{OB}(\text{help} \wedge \neg\text{crime})$ oddly entails $\text{OB}(\neg\text{crime})$, presupposing the crime that motivates the helping obligation.

3 MODALBENCH

3.1 PROBLEM GENERATION

Problems are generated algorithmically in four stages: (1) **Frame generation** with constrained randomness and post-hoc validation of system-specific properties; (2) **Valuation sampling** over 2–5 propositions; (3) **Formula selection** from tier-specific pools with axiom tagging; (4) **Ground truth computation** via recursive Kripke evaluation (Eqs. 1–2). No human annotation is needed. Label balance is enforced at exactly 50/50 True/False per cell; the "$\Box$ implies True" heuristic achieves only 47.1%, confirming resistance to shallow strategies.

3.2 DIFFICULTY TIERS

Tier 1 (Easy): Single modal operator, depth 1 ($\Box p$, $\Diamond(p \wedge q)$). **Tier 2 (Medium):** Nested modalities, depth 2 ($\Box\Diamond p$, $\Diamond(\Box p \wedge \Diamond q)$). **Tier 3 (Hard):** Axiom-interaction formulas and deontic paradoxes, each tagged with its axiom name (T, 4, 5, B, K, D, Ross, Chisholm, etc.). 23 distinct formula types and 5 paradox types.

3.3 DUAL-TRACK PRESENTATION

Each problem appears in two equivalent forms. The **formal track** states worlds, relations, and valuations explicitly with symbolic notation. The **natural language track** embeds the same Kripke model in a narrative: for alethic modalities, Alice explores rooms with one-way observation windows (accessibility = visibility); for deontic logic, jurisdictions exert regulatory influence. The gap between tracks reveals whether failures stem from modal reasoning itself or from the translation between formal and everyday modal language.

Figure 1 shows the same problem in both tracks. The formal track presents explicit set notation; the natural language track uses an observation-window narrative where "necessarily p" becomes "p is true in *all* rooms Alice can observe." Both tracks encode identical Kripke semantics and share the same computationally verified ground truth.

3.4 STATISTICS

ModalBench contains **3,000 problems** (1,500 unique $\times$ 2 tracks), 200 per system$\times$tier cell, frame sizes 2–7 worlds (mean 4.5), 23 axiom-tagged formula types, and 5 deontic paradox types.

4 EXPERIMENTAL SETUP

We evaluate three LLMs (Table 2) with three prompting strategies: (1) **Zero-shot:** "Answer True or False." (2) **Chain-of-thought (CoT):** "Think step by step." (3) **World-enumeration CoT:** Explicit instructions to enumerate accessible worlds, tabulate truth values, and aggregate—essentially externalizing the Kripke evaluation algorithm. All responses are generated with a 16k-token budget to

Formal Track (System T, Tier 2)

Kripke Model: Worlds = {Alpha, Beta, Gamma}. Accessibility: Alpha→Alpha, Alpha→Beta, Beta→Beta, Beta→Gamma, Gamma→Gamma.
Valuations: Alpha: p=True, q=False | Beta: p=False, q=True | Gamma: p=True, q=True.
Question: At world Alpha, is the following formula true or false? $\Box\Diamond q$
Answer: **True** (All accessible worlds satisfy $\Diamond q$: at Alpha, q holds in Beta; at Beta, q holds in Beta and Gamma.)

Natural Language Track (Same problem)

Alice is exploring a building. Some rooms have one-way observation windows into other rooms. Room Alpha can see itself and Room Beta. Room Beta can see itself and Room Gamma. Room Gamma can only see itself.

Facts: In Room Alpha: the switch is on, the lamp is **off**. In Room Beta: the switch is **off**, the lamp is on. In Room Gamma: the switch is on, the lamp is on.

Alice is currently in Room Alpha. Is the following statement true?
"*It is **necessarily** the case that the lamp being on is **possible**.*"
Answer: **True** (In every observable room, there exists a further observable room where the lamp is on.)

Figure 1: **Side-by-side prompt comparison.** The same Kripke model and formula presented in formal (top) and natural language (bottom) tracks. The NL version replaces $\Box$ with "necessarily," $\Diamond$ with "possible," and worlds with rooms connected by observation windows.

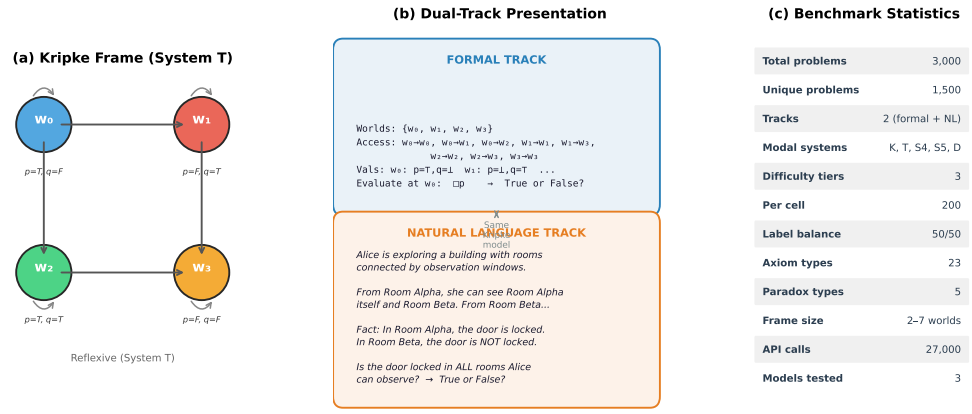

Figure 2: **The ModalBench Evaluation Framework.** Each of the 1,500 unique problems is grounded in computationally verified Kripke semantics (Left) and presented in both formal and natural language tracks (Center) to isolate pure modal reasoning capabilities from notation comprehension. The benchmark systematically scales across five modal systems and three difficulty tiers (Right).

avoid truncation-induced parse failures. Total: 27,000 inferences (3 models × 3 strategies × 3,000 problems).

5 Results

5.1 Main Results

Table 3 presents accuracy by model and modal system under CoT prompting. Gemini 2.5 Flash leads at **90.8%**, followed by Qwen3-235B (80.6%) and Llama 3.3 70B (72.8%). All models substantially exceed the 50% baseline, demonstrating genuine modal reasoning capability. System K (unconstrained accessibility) is consistently the hardest (65–72%), while S4 and T are the easiest. This suggests that structural regularity in the accessibility relation *helps* LLMs—the absence of constraints forces the model to reason without familiar patterns.

Table 2: Models evaluated. All accessed via API tiers.

Model	Type	Provider	Params	Parse Fail
Gemini 2.5 Flash	Reasoning	Google AI Studio	—	2.3%
Llama 3.3 70B	Standard	Groq	70B	7.8%
Qwen3-235B Instruct	MoE	Cerebras	235B	0.0%

Table 3: Accuracy (%) by model and modal system (CoT prompting, both tracks). System K is consistently hardest.

Model	D	K	S4	S5	T	Overall
Gemini 2.5 Flash	94.8	72.0	97.3	94.6	95.5	**90.8**
Llama 3.3 70B	71.7	65.6	73.8	75.8	77.3	72.8
Qwen3-235B Instruct	82.3	75.2	80.2	82.8	82.5	80.6

Tier degradation. Accuracy drops from 86.3% (Tier 1) to 79.1% (Tier 3), a 7.2 pp average decline (Table 4; ANOVA $p < 10^{-7}$ for Llama and Qwen3). Gemini shows minimal degradation (92.4% $\rightarrow$ 90.0%, $p = 0.11$), suggesting that its modal reasoning is robust even on axiom-interaction problems. Llama shows the steepest decline (79.2% $\rightarrow$ 67.7%, 11.5 pp).

5.2 Natural Language Advantage

A central question in logic evaluation is whether to present problems formally or in natural language. Table 5 reveals a striking finding: **two of three models perform significantly better on natural language presentations**. Qwen3-235B shows an 18.4 pp NL advantage ($t = -21.5$, $p < 10^{-99}$, Cohen's $d = -0.45$), and Llama shows a 4.7 pp NL advantage ($p < 10^{-6}$). Only Gemini is neutral ($+1.0$ pp, not significant).

This challenges the common assumption that symbolic notation should be easier for logical tasks. Our narrative descriptions use everyday modal language ("all rooms Alice can observe"), which appears to activate more effective reasoning pathways than abstract symbols ($\Box$, $\Diamond$). The effect is strongest for Qwen3-235B ($d = -0.45$, a medium effect size), suggesting that models with weaker formal-logic training benefit most from grounded, situated descriptions.

5.3 World-Enumeration as Scaffold

World-enumeration prompting externalizes the Kripke evaluation algorithm (Eqs. 1–2) as a step-by-step template. Rather than simply asking "think step by step," we instruct the model to:

1. *List* all worlds accessible from the evaluation world.
2. *Tabulate* the truth value of each sub-formula at each accessible world.
3. *Aggregate* using the correct quantifier: $\forall$ for $\Box$ ("all must hold") or $\exists$ for $\Diamond$ ("at least one must hold").
4. For nested formulas, *recurse* inside-out: evaluate the innermost sub-formula first, then use its results as input for the outer operator.

This template mirrors how a human logician would mechanically evaluate a Kripke formula. Table 6 shows this acts as a **scaffold for weaker models**: Llama improves by $+9.1$ pp and Qwen3 by $+10.7$ pp when moving from zero-shot to world-enumeration. However, Gemini—already at 95.9% zero-shot—slightly *declines* with structured prompting (-2.8 pp). All strategy effects are highly significant (ANOVA $p < 10^{-14}$).

Interpretation. When a model already possesses strong internal modal reasoning (Gemini), imposing an external evaluation template adds overhead without benefit. When a model has latent ability but struggles to organize its reasoning (Llama, Qwen3), the template provides the structure needed to succeed. This scaffold-vs-capability tradeoff has practical implications: the optimal prompting strategy depends on the model's baseline competence.

Table 4: Accuracy (%) by tier (CoT). Gemini's tier robustness contrasts with Llama's steep decline.

Model	Tier 1	Tier 2	Tier 3	ANOVA
Gemini 2.5 Flash	92.4	90.2	90.0	$F=2.2$, $p=0.11$
Llama 3.3 70B	79.2	71.6	67.7	$F=16.6$, $p<10^{-7}$***
Qwen3-235B Instruct	87.0	75.5	79.3	$F=22.3$, $p<10^{-10}$***

Table 5: Formal vs. natural language accuracy (%). Two of three models reason significantly *better* with narrative presentations. All strategies combined.

Model	Formal	NL	Gap	Significance
Gemini 2.5 Flash	93.8	92.8	$+1.0$ pp	$t=1.8$, $p=0.07$ (ns)
Llama 3.3 70B	72.9	77.5	-4.7 pp	$t=-4.9$, $p<10^{-6}$***
Qwen3-235B Instruct	68.4	86.8	-18.4 pp	$t=-21.5$, $p<10^{-99}$***

5.4 Implicit S5 Bias

We measure how often each model predicts True for axiom 5 formulas ($\Diamond P \to \Box\Diamond P$) by system, compared to ground truth (Table 7). Axiom 5 is valid only in S5; testing it in weaker systems reveals whether models have internalized the structural conditions for its validity.

In system K, all three models over-apply axiom 5: Qwen3 by $+56$ **pp**, Gemini by $+44$ pp, Llama by $+22$ pp. This means that when asked whether $\Diamond P \to \Box\Diamond P$ holds in a frame with no structural constraints, models systematically default to "True"—as if assuming universal accessibility. Qwen3 further over-applies in systems T ($+50$ pp) and S4 ($+22$ pp). Meanwhile, Llama under-applies axiom 5 in S5 (-27 pp), where it *is* valid. Models have not learned the relationship between frame properties and axiom validity.

5.5 Per-Axiom Diagnostics

Table 8 reveals which modal principles LLMs have internalized and which they have not.

Axiom 4 ($\Box P \to \Box\Box P$, positive introspection) is near chance for Llama (48.5%) but substantially above chance for Gemini (76.3%) and Qwen3 (65.8%). **Axiom K** (distribution, always valid) is perfectly recognized by Gemini and Qwen3. The **dual** ($\Box p \leftrightarrow \neg\Diamond\neg p$) is correctly handled by Gemini (97.5%) and Qwen3 (95.0%) but not Llama (56.4%), indicating that Llama has not internalized the necessity-possibility equivalence. The **converse of T** ($\Diamond p \to p$, *not valid*) is correctly rejected by Gemini 91.9% of the time.

5.6 Deontic Paradox Results

Ross's Paradox is universally solved (100%). FO/PE Contradiction ($\text{FO}(p) \wedge \text{PE}(p)$, should always be False) is now well-handled (89–98%). The Gentle Murderer (contrary-to-duty obligations) remains hardest, with all models below 79%, confirming that reasoning about violated obligations is a genuine frontier.

5.7 Frame Size and Nesting Depth

When models are given sufficient generation budget (8k–16k tokens), accuracy is *flat* with respect to frame size (Figure 5a): Gemini averages $\sim$90%, Qwen3 $\sim$80%, and Llama $\sim$72% across frame sizes from 2 to 7 worlds, with no significant monotone trend (Spearman $|\rho| \leq 0.26$, all $p > 0.6$). This suggests that Kripke model size is not a bottleneck for modern LLMs when responses can be produced in full.

Nesting depth does affect accuracy (Figure 5b): Llama drops from 76.6% (depth 1) to 67.6% (depth 2), and Qwen3 drops from 86.3% to 72.5%. Gemini is less affected (92.4% $\to$ 88.7%).

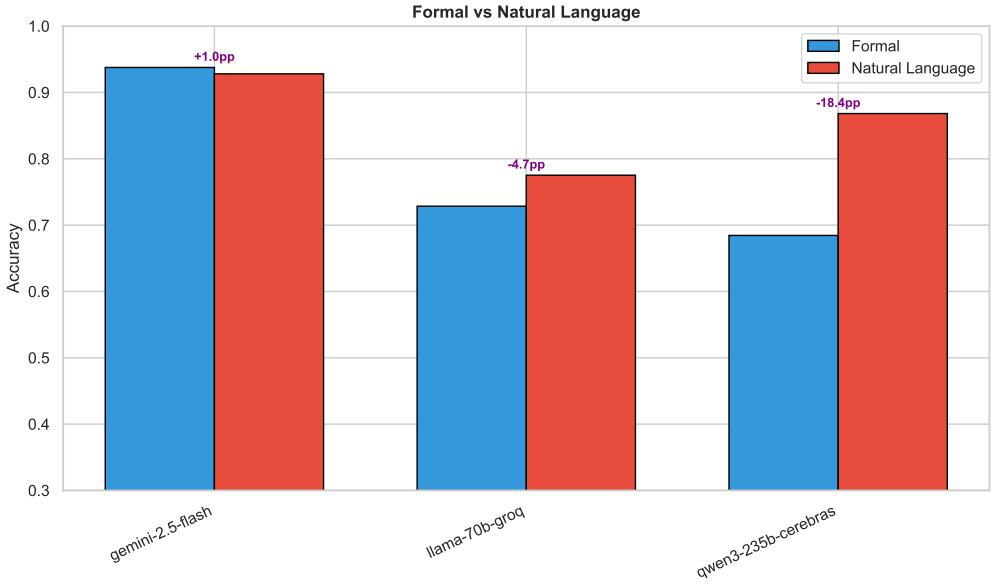

Figure 3: **The Natural Language Advantage.** A direct comparison of model accuracy between formal symbolic notation and natural language narrative tracks. Challenging the historical assumption that mathematical logic requires symbolic representation, both Qwen3-235B (+18.4 percentage points) and Llama 3.3 70B (+4.7 percentage points) demonstrate highly significant performance improvements when reasoning through grounded narratives rather than abstract Kripke semantics. Gemini 2.5 Flash remains relatively neutral across both modalities.

Table 6: Accuracy (%) by prompting strategy. World-enumeration scaffolds weaker models but is unnecessary for the strongest. All effects significant ($p < 10^{-14}$).

Model	Zero-shot	CoT	World-enum	Best
Gemini 2.5 Flash	**95.9**	90.8	93.1	Zero-shot
Llama 3.3 70B	72.3	72.8	**81.4**	World-enum
Qwen3-235B Instruct	70.8	80.6	**81.5**	World-enum

Depth 2 formulas ($\Box\Diamond p$, $\Diamond\Box p$) require recursive world quantification, doubling the reasoning burden.

6 Related Work

Logic reasoning benchmarks. The landscape of LLM logic evaluation is rich but restricted to classical logics. FOLIO (Han et al., 2024) provides expert-annotated first-order natural language inference problems with verified FOL translations. LogiQA (Liu et al., 2020) draws from standardized exam questions testing logical comprehension. ProofWriter (Tafjord et al., 2021) generates multi-step deductive proofs over synthetic rule bases, while PrOntoQA (Saparov & He, 2023) uses ontologies to systematically test chain-of-thought reasoning on syllogisms. LogicBench (Parmar et al., 2024) provides broader coverage across multiple logic types including propositional, predicate, and non-monotonic reasoning. Despite this richness, *none* of these benchmarks address modal or deontic logic, nor do they employ Kripke-semantic ground truth. ModalBench occupies a distinct niche: it is the only benchmark where truth values are computed by recursive evaluation on explicit possible-worlds models, rather than derived from natural language annotations or logical entailment rules.

Table 7: Implicit S5 bias: predicted True rate minus ground truth True rate for axiom 5. Positive values indicate over-application. All three models over-apply axiom 5 in system K, where it is not valid.

Model	K	S4	S5	T
Gemini 2.5 Flash	+0.44	+0.00	+0.00	+0.00
Llama 3.3 70B	+0.22	+0.00	−0.27	+0.00
Qwen3-235B Instruct	+0.56	+0.22	+0.00	+0.50

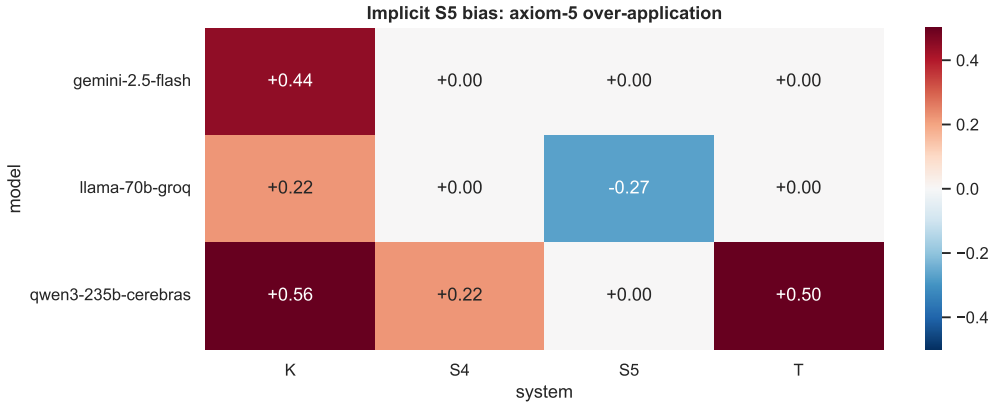

Figure 4: **Implicit S5 Bias and Axiom Hallucination.** A heatmap representing the rate at which models predict 'True' for Axiom 5 (the Euclidean property) compared to the ground-truth validity. Red indicates over-application; blue indicates under-application. All models systematically over-apply Axiom 5 in System K (where no structural constraints exist), with Qwen3-235B over-applying by up to +56 percentage points. This reveals a fundamental reasoning flaw: models routinely hallucinate universal accessibility (S5 semantics) even when explicitly prompted to reason within weaker, unconstrained modal systems.

Implicit modal reasoning in NLP. Several NLP tasks implicitly require modal reasoning without formalizing it. Theory-of-mind benchmarks (Sap et al., 2022) test whether models can track epistemic states ("Alice *believes* that P")—a form of epistemic modal reasoning. The Sally-Anne test and BigToM evaluate belief attribution, which corresponds to reasoning in epistemic logic S5. However, these tasks lack formal ground truth: when a model fails, it is impossible to determine whether the failure stems from a modal reasoning error (misapplying $\Box/\Diamond$) or from a natural language understanding error (misinterpreting "believes"). Our dual-track design explicitly disentangles these failure modes—and reveals that NL understanding is *not* the bottleneck; most models actually perform better with natural language.

Legal and deontic reasoning. LegalBench (Guha et al., 2024) evaluates legal reasoning but frames tasks as NL comprehension rather than formal deontic evaluation. ModalBench complements it by testing the *structural properties* of obligation, permission, and prohibition—including paradoxes from their formal definitions—rather than their application in natural-language legal contexts.

Logical consistency and calibration. Cheng et al. (2025) provide a comprehensive survey of LLM logical reasoning, explicitly identifying "extending to modal logic" as an open direction. Ghosh et al. (2024) study logical consistency under propositional operators (negation, conjunction, disjunction) but not modal ones. Our implicit S5 bias finding extends consistency analysis to the modal setting: we show that models are *systematically* inconsistent across modal systems, over-applying axioms from S5 even when explicitly evaluated in weaker systems. This is a new form of logical inconsistency that has no analog in propositional benchmarks.

Table 8: Accuracy (%) on selected axioms (CoT, both tracks combined). Axiom K is well-recognized; axiom 4 remains hard for Llama. Full results (23 axiom types) in Appendix A.

Axiom	Formula	Gemini	Llama	Qwen3
K (distrib.)†	$\Box(p \to q) \to (\Box p \to \Box q)$	**100.0**	58.3	**100.0**
T (reflex.)	$\Box p \to p$	82.1	60.7	**85.7**
4 (transit.)	$\Box p \to \Box\Box p$	**76.3**	48.5	65.8
5 (Euclid.)	$\Diamond p \to \Box\Diamond p$	**87.9**	57.6	63.6
B (symmetry)	$p \to \Box\Diamond p$	**78.8**	53.1	61.5
Dual†	$\Box p \to \neg\Diamond\neg p$	**97.5**	56.4	95.0
Conv. T (invalid)	$\Diamond p \to p$	**91.9**	73.8	74.2

Table 9: Deontic paradox accuracy (%, CoT). Gentle Murderer (contrary-to-duty) remains the hardest.

Paradox	Gemini	Llama	Qwen3
Ross's Paradox	100.0	100.0	100.0
Chisholm's Paradox	100.0	55.6	90.0
FO/PE Contradiction	98.4	88.7	96.8
Gentle Murderer	78.9	63.2	65.8
Good Samaritan	66.7	83.3	100.0

Neuro-symbolic approaches. Logic-LM (Pan et al., 2023) and LINC (Olausson et al., 2023) combine LLMs with symbolic solvers for first-order logic. In these pipelines, the LLM translates natural language to formal logic, and a solver handles the deduction. ModalBench provides the formal structures needed to extend such pipelines to modal logic: an LLM could extract the Kripke model from a natural-language description, and a Kripke evaluator could compute the formula's truth value exactly. Our natural language advantage finding (§5.2) suggests that the extraction step—NL to Kripke structure—may be easier than the evaluation step for current models.

7 Limitations and Future Work

We evaluate three models; including open-weight models for hidden-state probing—to test whether models develop internal representations of accessibility relations—is an important direction. Our NL track uses controlled narrative scenarios; real-world modal language ("must," "might," "should") is more ambiguous and context-dependent. We do not address multi-agent epistemic logic, dynamic logic, or temporal logic. Llama's 7.8% parse failure rate suggests room for more robust answer extraction. Future work should explore adaptive prompting that selects the strategy based on model capability, and neuro-symbolic pipelines where an LLM extracts Kripke structure and a solver evaluates the formula exactly.

8 Conclusion

ModalBench reveals that LLMs can reason about modal logic at 73–91% accuracy, but with systematic, interpretable failure patterns. Most strikingly, two of three models reason *better* in natural language than in formal notation—with Qwen3-235B showing an 18.4 pp NL advantage—challenging the assumption that symbolic presentation is inherently easier for logical tasks. World-enumeration prompting acts as a scaffold for weaker models (+9–11 pp) but is unnecessary for the strongest model (95.9% zero-shot). All models exhibit implicit S5 bias, over-applying the Euclidean axiom by up to +56 pp in system K. System K (unconstrained accessibility) is consistently the hardest, and contrary-to-duty deontic reasoning remains a genuine frontier. We release ModalBench to support research on modal reasoning as a critical capability for trustworthy LLMs.

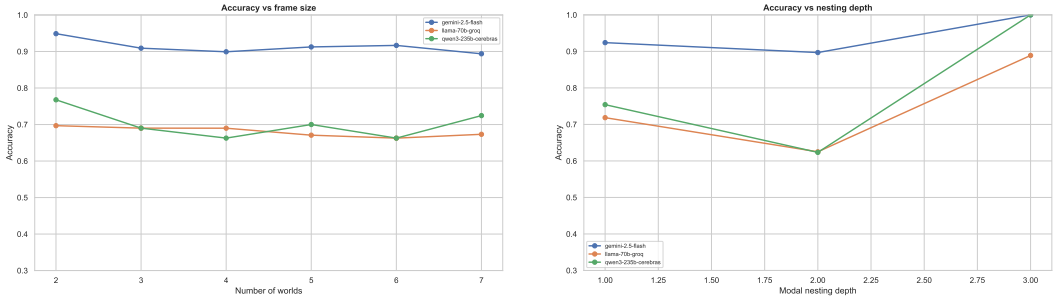

Figure 5: **Scaling Behavior: Breadth vs. Depth.** **(a)** Accuracy across frame size (number of worlds) remains remarkably stable for all models (Spearman $|\rho| \leq 0.26$, $p > 0.6$), demonstrating that combinatorial breadth does not degrade performance once complete reasoning chains are established. **(b)** Conversely, modal nesting depth drives a distinct decline in accuracy, particularly for Llama 3.3 70B and Qwen3-235B. This reveals that recursive logical tracking—not the sheer size of the Kripke frame—is the primary computational bottleneck for LLM modal reasoning.

ACKNOWLEDGMENTS

We thank the anonymous reviewers of the ICLR 2026 Workshop on Logical Reasoning of Large Language Models for their insightful and constructive feedback, which significantly strengthened our parsing methodology and the depth of our diagnostic analysis. We also extend our deepest gratitude to Mr. Shiraz Zaman and Mr. Martin Liu, the founders of NAND AI. The foundational experience in LLM evaluation gained while working at NAND AI, along with their constant guidance, motivation, and mentorship, was instrumental in shaping the direction of this research.

REPRODUCIBILITY STATEMENT

ModalBench is fully algorithmic and deterministic given a random seed. All models were accessed via API tiers. Code and data: `https://github.com/mujtabahasan/modalbench`.

REFERENCES

Brian F Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

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

Bishwamittra Ghosh et al. Logical consistency of large language models in fact-checking. In *arXiv preprint arXiv:2412.16100*, 2024.

Neel Guha et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 2024.

Simeng Han et al. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

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

Jian Liu et al. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Paul McNamara and Frederik Van De Putte. Deontic logic. *Stanford Encyclopedia of Philosophy*, 2021.

Theo X Olausson et al. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of EMNLP*, 2023.

Liangming Pan et al. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of EMNLP*, 2023.

Mihir Parmar et al. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of ACL*, 2024.

Maarten Sap et al. Neural theory-of-mind? on the limits of social intelligence in large LMs. *arXiv preprint arXiv:2210.13312*, 2022.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *Proceedings of ICLR*, 2023.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of ACL*, 2021.

A Full Per-Axiom Results

Table 10: Complete per-axiom accuracy (%, CoT, both tracks combined). † = valid in all normal modal logics.

Tag	Description	Gemini	Llama	Qwen3
K†	Distribution	**100.0**	58.3	**100.0**
T	Reflexivity	82.1	60.7	**85.7**
4	Transitivity	**76.3**	48.5	65.8
5	Euclidean	**87.9**	57.6	63.6
B	Symmetry	**78.8**	53.1	61.5
dual†	$\Box p \to \neg\Diamond\neg p$	**97.5**	56.4	95.0
dual_rev†	$\neg\Diamond\neg p \to \Box p$	**88.0**	34.8	80.8
dist†	$\Box(p\wedge q) \to \Box p\wedge\Box q$	**96.4**	53.6	75.0
conv_T	$\Diamond p \to p$ (invalid)	**91.9**	73.8	74.2
Barcan	$\Diamond\Box p \to \Box\Diamond p$	**87.5**	61.3	78.1
McKinsey	$\Box\Diamond p \to \Diamond\Box p$	**83.3**	46.2	50.0
triple	$\Box\Box\Box p \to \Box p$	**94.4**	78.8	**100.0**
D	$\mathrm{OB}(p) \to \mathrm{PE}(p)$	**100.0**	70.0	**100.0**
Ross	$\mathrm{OB}(p) \to \mathrm{OB}(p\vee q)$	**100.0**	**100.0**	**100.0**
Chisholm	Contrary-to-duty	**100.0**	55.6	90.0
Gentle	Gentle Murderer	**78.9**	63.2	65.8
FO_PE_contra	Contradiction	**98.4**	88.7	96.8
OB_4	$\mathrm{OB}(\mathrm{OB}(p))\to\mathrm{OB}(p)$	**100.0**	75.0	87.5
PE_dist	$\mathrm{PE}(p\wedge q)\to\mathrm{PE}(p)\wedge\mathrm{PE}(q)$	**93.8**	**87.5**	**93.8**
ought_can	$\mathrm{OB}(p) \to \Diamond p$	**100.0**	**87.5**	93.8
no_conflict	$\neg(\mathrm{OB}(p)\wedge\mathrm{OB}(\neg p))$	**100.0**	**100.0**	**100.0**

B World-Enumeration Prompt Template

The following is the exact world-enumeration prompt template appended to each problem. This template externalizes the recursive Kripke evaluation algorithm as natural-language instructions:

> *To solve this problem, follow these steps exactly:*
>
> *Step 1: Identify the evaluation world and list ALL worlds accessible from it.*
>
> *Step 2: For the outermost modal operator, determine the type:*
>
> - *If □ (necessarily / obligatory): the sub-formula must be true at ALL accessible worlds.*
> - *If ◇ (possibly / permitted): the sub-formula must be true at at least ONE accessible world.*
>
> *Step 3: Evaluate the sub-formula at each accessible world. If the sub-formula itself contains modal operators, recurse: for each world, determine its accessible worlds and repeat.*
>
> *Step 4: Combine your results using the quantifier from Step 2.*
>
> *Step 5: State your final answer as exactly "True" or "False."*

This template mirrors how Kripke evaluation works mathematically: enumerate, evaluate, aggregate. Models that struggle in zero-shot settings often fail at the organizational step—forgetting to check all accessible worlds or aggregating with the wrong quantifier. The template eliminates these organizational failures.

C DEONTIC PROBLEM EXAMPLE

The following is a sample deontic problem from Tier 3 (Ross's Paradox):

> **System D (Deontic).** Jurisdictions: {Alpha, Beta, Gamma}. Regulatory influence: Alpha influences Alpha and Beta; Beta influences Beta and Gamma; Gamma influences only itself.
>
> *Regulations:* In Alpha: voting is mandatory, recycling is optional. In Beta: voting is optional, recycling is mandatory. In Gamma: voting is mandatory, recycling is mandatory.
>
> Under jurisdiction Alpha, is the following statement true?
>
> "*If recycling is **obligatory**, then recycling-or-littering is **obligatory**.*"
>
> *(Formula:* $\text{OB}(\textit{recycling}) \rightarrow \text{OB}(\textit{recycling} \vee \textit{littering})$*)*
>
> *Answer:* **True.** The antecedent is $\text{OB}(\text{recycling})$: recycling must hold in all influenced jurisdictions (Alpha and Beta). Recycling is optional at Alpha, so $\text{OB}(\text{recycling})$ is False. Since the antecedent is False, the implication is vacuously True. This captures Ross's Paradox: the formula is valid in SDL regardless of context, because $\text{OB}(p) \rightarrow \text{OB}(p \vee q)$ is a theorem.

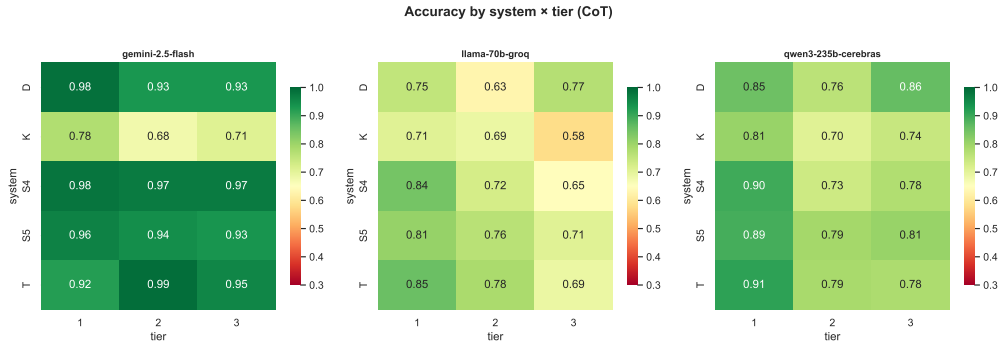

Figure 6: **Performance Degradation Across Systems and Tiers.** A granular heatmap of Chain-of-Thought (CoT) accuracy across all five modal systems and three difficulty tiers. While Gemini 2.5 Flash exhibits uniform robustness across the matrix, Llama 3.3 70B and Qwen3-235B reveal significant, systematic vulnerabilities. Specifically, both open-weight models struggle heavily with System K—where the absence of accessibility constraints forces pure, unguided logical evaluation—and Tier 3, which requires complex axiom-interaction tracking. (Darker colors indicate higher accuracy).