# ChaosBench-Logic v2: Evaluating LLM Logical Reasoning over Dynamical Systems at Scale

Noel Thomas
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
noel.thomas@mbzuai.ac.ae

## Abstract

Standard accuracy on binary reasoning benchmarks hides critical failure modes: prior collapse, inconsistency under paraphrase, and inability to reason about parameter-dependent dynamics. We present ChaosBench-Logic v2, a 40,886-question benchmark over 165 dynamical systems with 27 FOL predicates and 78 axiom edges, together with CARE (Calibration- and Adversarial-Robust Evaluation), a protocol that surfaces these pathologies. Evaluating 14 models, we find that regime transition reasoning remains near-random (MCC = 0.05) even for frontier models, while FOL deduction with given premises reaches MCC = 0.52; per-family decomposition shows the proprietary advantage concentrates on cross-indicator (+0.40) and consistency tasks, while open-source Qwen 2.5-32B dominates indicator diagnostics (0.91 vs. 0.45). Two models exhibit negative MCC on bifurcation questions, confirmed as systematic anti-correlation via confusion matrix analysis.

## 1 Introduction

Large language models achieve strong performance on mathematical and logical reasoning benchmarks (Wei et al., 2022; Cobbe et al., 2021), yet their capacity for logically consistent reasoning over scientific domains remains poorly understood. Dynamical systems present a particularly demanding testbed: chaos is deterministic but not random, exhibits sensitive dependence on initial conditions, and requires positive Lyapunov exponents (Strogatz, 2018). These formal distinctions must be maintained across multi-step inferences, a requirement that probes deeper than pattern matching.

Existing benchmarks target mathematical problem-solving (Hendrycks et al., 2021; Cobbe et al., 2021), scientific QA (Clark et al., 2018; Wang et al., 2023a), propositional logic (Liu et al., 2020; Han et al., 2022), or synthetic FOL reasoning (Saparov & He, 2023), but none combine (i) a domain-specific FOL ontology with (ii) ground-truth labels derived from axiom entailment over (iii) real scientific systems at scale.

We introduce ChaosBench-Logic v2, a 66× scale-up of v1 (Thomas, 2026): from 621 to 40,886 questions, 27 to 165 systems, and 7 to 11 task families. Our contributions:

1. Benchmark. 40,886 questions, 11 task families, 27 predicates, 78 FOL axiom edges, 165 dynamical systems (135 from dysts (Gilpin, 2021)).
2. CARE protocol. A calibration- and robustness-aware evaluation framework (MCC, macro-family MCC, calibration diagnostics, consistency, coverage) that exposes failure modes hidden by accuracy.
3. Diagnostic findings. A knowledge-type boundary between rule-following and parameter-dependent reasoning, per-family decomposition of the proprietary–OSS gap, and systematic prediction biases including negative MCC.

## 2 Related Work

LLM reasoning benchmarks. GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) test mathematical reasoning; ARC (Clark et al., 2018) tests science QA; LogiQA (Liu et al., 2020) and FOLIO (Han et al., 2022) test logical reasoning; BIG-Bench (Srivastava et al., 2023) includes some logical tasks. PrOntoQA (Saparov & He, 2023) is closest to our work: it tests compositional FOL reasoning over synthetic ontologies, finding that LLMs struggle with longer inference chains. LogicBench (Parmar et al., 2024) evaluates 25 logical reasoning patterns and confirms failures on complex reasoning with negation. Our benchmark differs by grounding FOL axioms in a real scientific domain where ground-truth labels derive from physical properties rather than synthetic constructions.

Consistency and robustness. TruthfulQA (Lin et al., 2022) measures truthfulness; self-consistency decoding (Wang et al., 2023b) improves CoT reliability; ReClor (Yu et al., 2020) tests logical reading comprehension. Our consistency_paraphrase and perturbation families extend these ideas to a scientific domain with formal ground truth.

Scientific reasoning and dynamical systems. SciBench (Wang et al., 2023a) evaluates college-level scientific problem-solving. The dysts library (Gilpin, 2021) provides 135 standardized dynamical systems originally designed for forecasting benchmarks; we build on it for system diversity, preserving provenance and verifying predicate annotations against our axiom system. ChaosBench-Logic v1 (Thomas, 2026) introduced a 621-question benchmark; we scale by two orders of magnitude.

## 3 Benchmark Design

### 3.1 Ontology

The benchmark is grounded in a first-order logic ontology of 27 unary predicates in three tiers: 11 core predicates characterizing dynamical regimes (Chaotic, Deterministic, PosLyap, Sensitive, StrangeAttr, PointUnpredictable, StatPredictable, QuasiPeriodic, Random, FixedPointAttr, Periodic), 4 topological predicates (Dissipative, Bounded, Mixing, Ergodic), and 12 structural predicates (HyperChaotic, Conservative, ContinuousTime, DiscreteTime, etc.).

These predicates are connected by 78 directed axiom edges (31 implication, 47 exclusion), enabling reasoning chains of up to 5–6 hops. For example, the Chaotic predicate entails:

$$\begin{aligned}\forall s:\ \mathrm{Chaotic}(s) \Rightarrow\ & \mathrm{Deterministic}(s) \wedge \mathrm{PosLyap}(s) \wedge \mathrm{Sensitive}(s) \wedge \mathrm{Mixing}(s) \\ & \wedge\ \neg\mathrm{Random}(s) \wedge \neg\mathrm{Periodic}(s) \wedge \neg\mathrm{QuasiPeriodic}(s) \end{aligned} \tag{1}$$

The full specification is in Appendix F.

### 3.2 Systems

The benchmark covers 165 dynamical systems:[1] 30 manually curated (Lorenz-63 (Lorenz, 1963), Rössler, Hénon, logistic map, Brusselator, Ornstein-Uhlenbeck, etc.) and 135 from the dysts library (Gilpin, 2021). Each system carries ground-truth values for all 27 predicates, verified against the axiom system.

### 3.3 Task Families

Multi-hop questions chain 2–6 steps through the axiom graph (e.g., "FluidTrampoline is strongly mixing ⇒ weakly mixing ⇒ ergodic ⇒ bounded. Is it bounded?"). Regime transition questions require specific bifurcation thresholds (e.g., "At $\alpha$=15.6, is Chua's circuit chaotic?"). FOL inference questions present premises for deductive conclusions. Adversarial questions include misleading premises irrelevant to the answer. Indicator diagnostic

[1]Regime transition and FOL inference families additionally use synthetic parameterizations not counted in this total.

Table 1: Dataset composition by task family, ordered by difficulty. $N < 100$ families are interpreted qualitatively.

| Family | N | Description |
|---|---|---|
| regime_transition | 68 | Bifurcation-dependent behavior |
| cross_indicator | 67 | Multi-indicator reasoning |
| consistency_paraphrase | 4,139 | Linguistic variation stability |
| perturbation | 1,994 | Parameter perturbation robustness |
| atomic | 25,307 | Single-predicate queries |
| adversarial_nearmiss | 478 | Near-miss misconceptions |
| adversarial_misleading | 500 | Misleading cue probes |
| multi_hop | 6,000 | 2–6 step inference chains |
| fol_inference | 1,758 | FOL deduction from premises |
| indicator_diagnostic | 530 | Chaos indicator interpretation |
| extended_systems | 45 | Factual recall (ceiling check) |
| Total | 40,886 | |

questions require interpreting numerical chaos indicators (0-1 test (Gottwald & Melbourne, 2009), permutation entropy (Bandt & Pompe, 2002), MEGNO (Cincotta & Simó, 2000)). Representative examples with model predictions are in Appendix E.

### 3.4 Evaluation Protocol: CARE

Standard accuracy on binary classification benchmarks can be misleading. Since Acc = $p \cdot \text{TPR} + (1-p) \cdot \text{TNR}$ where $p$ is the TRUE prevalence, a model with high TNR but low TPR inflates accuracy by exploiting class priors. LLaMA 3.1-8B illustrates this: its TPR = 0.32 and TNR = 0.88 yield 60.2% accuracy (vs. 50.5% for always-FALSE), but MCC = 0.24 and balanced accuracy = 0.60 correctly signal near-chance performance. To surface such pathologies, we propose CARE (Calibration- and Adversarial-Robust Evaluation), a protocol for reasoning benchmarks.

CARE reports five diagnostics:

1. MCC (primary) (Matthews, 1975; Chicco & Jurman, 2020): penalizes prior collapse and is invariant to class balance. Ranges from $-1$ (anti-correlation) through 0 (random) to $+1$ (perfect):

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

2. Macro-family MCC: mean MCC across task families, preventing dominant families (atomic: 62% of questions) from masking failures on hard families.
3. Calibration: predicted TRUE rate vs. ground-truth TRUE rate. Flags prior collapse (e.g., LLaMA 3.1-8B predicts TRUE only 21.6% vs. 49.5% ground truth).
4. Consistency: MCC on consistency_paraphrase and perturbation families, measuring whether predictions survive surface-level variation.
5. Coverage: invalid rate (unparseable responses counted as incorrect) and per-family coverage, flagging instruction-following failures.

Table 2 demonstrates CARE on four models. Accuracy ranks LLaMA 3.1-8B at 60.2% (above chance), but CARE flags prior collapse (21.6% predicted TRUE vs. 49.5% ground truth) and asymmetric recall (TPR = 0.32, TNR = 0.88): the model achieves accuracy by defaulting to FALSE, not by reasoning. Mistral-7B's 61.3% accuracy masks a 1.1% invalid rate and consistency MCC below 0.30. Only Claude Sonnet 4.6 triggers no CARE flags.

All standard models use temperature = 0; reasoning models (o3-mini, GPT-5.2) do not accept a temperature parameter and are deterministic by design (Appendix I). Most models receive max_tokens = 16; reasoning models and Gemini use 1024 because their architectures consume output tokens for internal reasoning.

Table 2: CARE diagnostics for four models. Flags: pc = prior collapse (|pred TRUE% − 49.5| > 10); ar = asymmetric recall (|TPR − TNR| > 0.3); cv = coverage (>0.5% invalid); ic = inconsistent (consistency MCC < 0.30).

| Model | Acc | MCC | TPR | TNR | Pred T% | Flags |
|---|---|---|---|---|---|---|
| Claude Sonnet 4.6 | 79.8 | 0.601 | 0.72 | 0.87 | 42.5 | (none) |
| Qwen 2.5-32B | 73.8 | 0.478 | 0.68 | 0.79 | 44.4 | ic |
| LLaMA 3.1-8B | 60.2 | 0.240 | 0.32 | 0.88 | 21.6 | pc, ar, ic |
| Mistral-7B | 61.3 | 0.228 | 0.67 | 0.55 | 56.2 | cv, ic |

Table 3: Full canonical (N = 40,886), ranked by MCC.

| Model | Type | Bal. Acc | MCC |
|---|---|---|---|
| Claude Sonnet 4.6 | Prop. | 0.797 | 0.601 |
| GPT-5.2 | Prop. | 0.747 | 0.509 |
| Qwen 2.5-32B | OSS | 0.738 | 0.478 |
| DeepSeek-Chat | Prop. | 0.724 | 0.469 |
| Gemini 2.5 Flash | Prop. | 0.718 | 0.458 |
| GPT-4o | Prop. | 0.721 | 0.450 |
| Qwen 2.5-14B | OSS | 0.711 | 0.426 |
| LLaMA 3.3-70B | OSS | 0.681 | 0.373 |
| LLaMA 3.1-8B | OSS | 0.599 | 0.240 |
| Mistral-7B | OSS | 0.613 | 0.228 |

# 4 Experiments

We evaluate 14 models: 7 proprietary (Claude Sonnet 4.6, GPT-5.2, GPT-4o, GPT-4o-mini, o3-mini, Gemini 2.5 Flash, DeepSeek-Chat) and 7 open-source (Qwen 2.5-{7B,14B,32B}, LLaMA 3.3-70B, LLaMA 3.1-8B, Gemma2-9B, Mistral-7B) served via Ollama. Ten models complete the full dataset (N = 40,886); four use subsets (5k or 1k) due to compute constraints, reported in a separate table to avoid cross-$N$ ranking.

# 5 Results

## 5.1 Overall Performance

Claude Sonnet 4.6 leads with MCC = 0.601 (Table 3). The proprietary–OSS gap is 0.12 MCC, but Qwen 2.5-32B (0.478) outperforms GPT-4o (0.450) and Gemini 2.5 Flash (0.458). Subset evaluations (o3-mini MCC = 0.608 on 5k; Gemma2-9B 0.280 on 5k; GPT-4o-mini 0.272 on 1k; Qwen 2.5-7B 0.268 on 1k) are reported in Appendix A.

## 5.2 Task Family Hardness

Figure 1 ranks families by mean MCC. Easy (MCC > 0.5): extended_systems (0.81, ceiling effect), indicator_diagnostic (0.59), fol_inference (0.52). Medium (0.25–0.5): multi_hop (0.48), adversarial families (0.44–0.45), atomic (0.32). Hard (< 0.25): perturbation (0.26), consistency_paraphrase (0.25), cross_indicator (0.18), regime_transition (0.05). Regime transition is near-random for all models: these questions require specific bifurcation thresholds (e.g., logistic map at $r \approx 3.57$) not recoverable from logical rules.

## 5.3 Per-Model Family Analysis

Figure 2 reveals that family-level performance is not monotonic with overall MCC. Qwen 2.5-32B achieves MCC = 0.91 on indicator_diagnostic (exceeding GPT-4o at 0.89 and Claude Sonnet at 0.45), while Claude Sonnet leads on multi_hop (0.64) and atomic (0.62). LLaMA 3.1-8B scores perfectly on extended_systems (45 factual-recall questions, ceiling effect) but near-zero on cross_indicator.

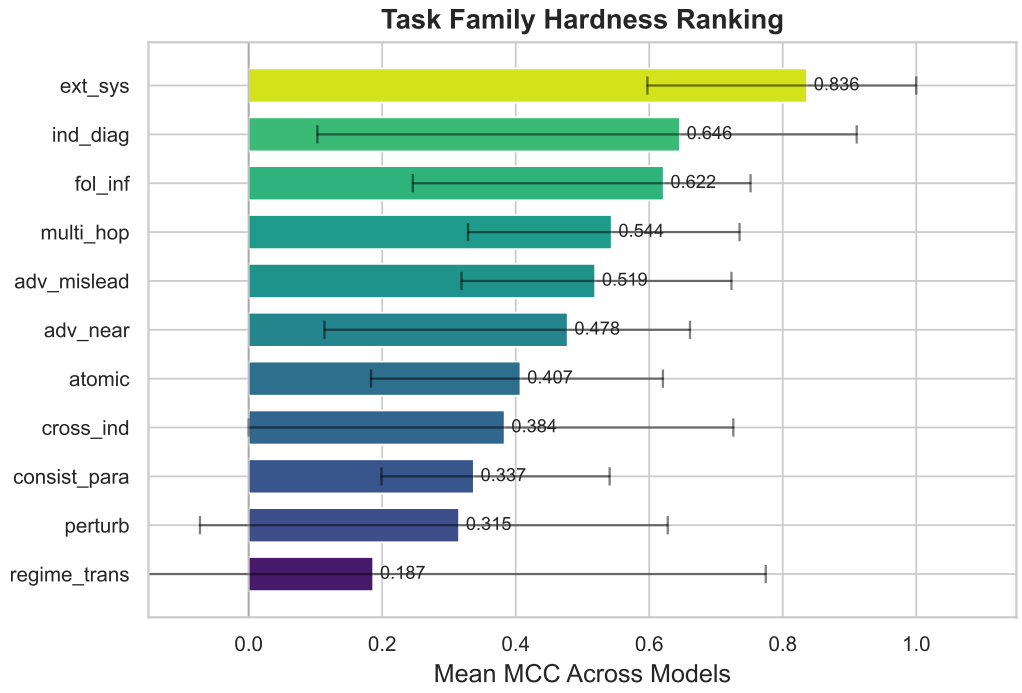


Figure 1: Mean MCC by task family across 10 full-canonical models. Error bars: min–max.

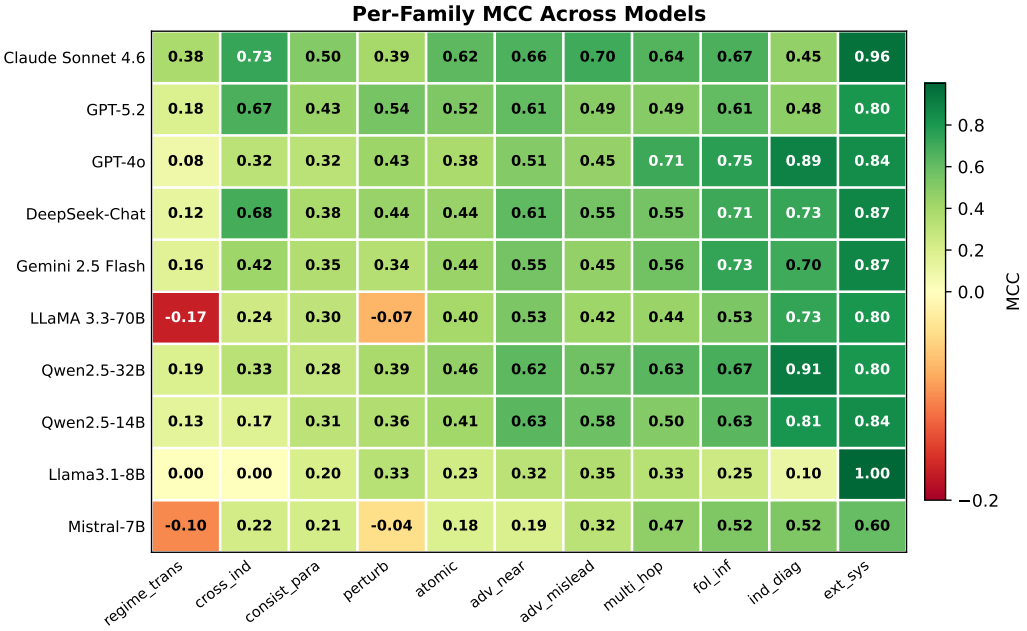


Figure 2: Per-family MCC for 10 models. Families ordered by hardness (left = hardest). Red cells indicate negative MCC (anti-correlation).

Two models produce negative MCC on regime_transition: LLaMA 3.3-70B ($-0.17$; TP = 9, FP = 17, TN = 20, FN = 22; balanced accuracy 0.42) and Mistral-7B ($-0.10$). Negative

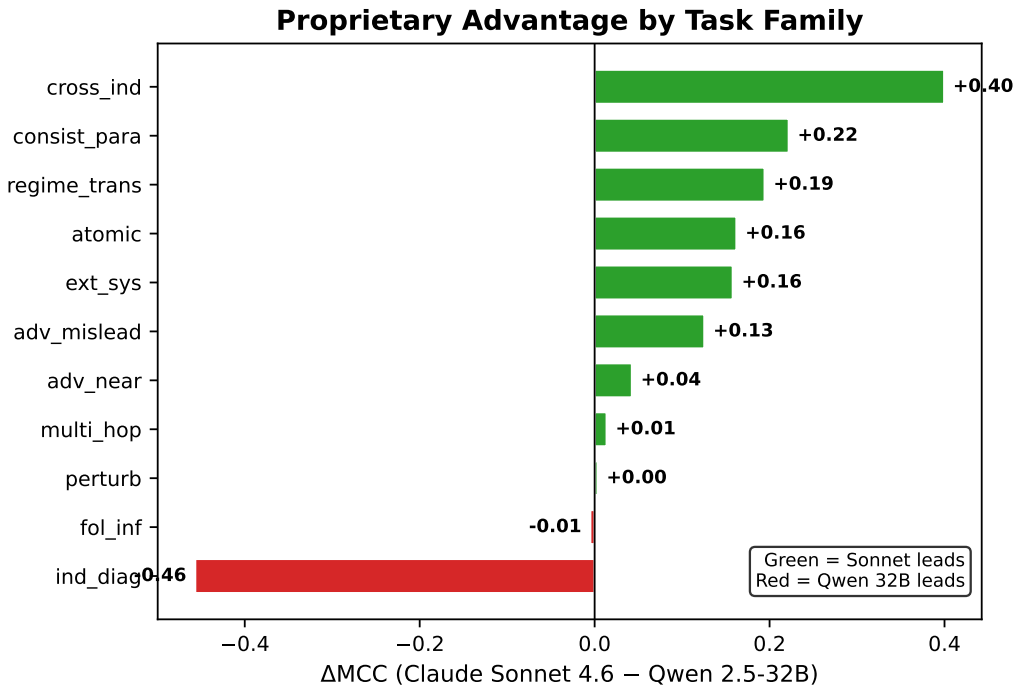


Figure 3: Per-family ΔMCC (Claude Sonnet 4.6 − Qwen 2.5-32B). Green: Sonnet leads. Red: Qwen leads.

MCC means systematic anti-correlation: these models have learned heuristics that are reliably wrong on bifurcation questions. Confusion matrices are in Appendix G.

### 5.4 Where the Proprietary Advantage Concentrates

The 0.12 overall gap is not uniform (Figure 3). Sonnet's advantages concentrate on cross_indicator ($\Delta = +0.40$), consistency_paraphrase (+0.22), and regime_transition (+0.19): families requiring integration of quantitative signals or robustness to surface variation. Near-parity on perturbation (0.00), multi_hop (+0.01), fol_inference (−0.01). Qwen 2.5-32B leads decisively on indicator_diagnostic (−0.46).

### 5.5 Prediction Bias

The ground-truth TRUE rate is 49.5%. Most models predict TRUE 42–50%, but LLaMA 3.1-8B predicts TRUE only 21.6% (TNR = 0.88, TPR = 0.32), explaining its low MCC despite moderate balanced accuracy. Mistral-7B shows the opposite: 56.2% predicted TRUE (Figure 4).

## 6 Discussion

A knowledge-type boundary. The central finding is a dissociation between two types of reasoning. FOL inference (MCC = 0.52) tests whether models can apply deductive rules when premises are explicitly stated; regime transition (MCC = 0.05) tests whether they can supply numerical premises themselves (e.g., the logistic map transitions to chaos at $r \approx 3.57$). This is not a scaling problem: within the Qwen 2.5 family, increasing from 7B to 32B parameters improves multi-hop and FOL inference but leaves regime transition near-random (Appendix B). The gap identifies a precise boundary between what LLMs can

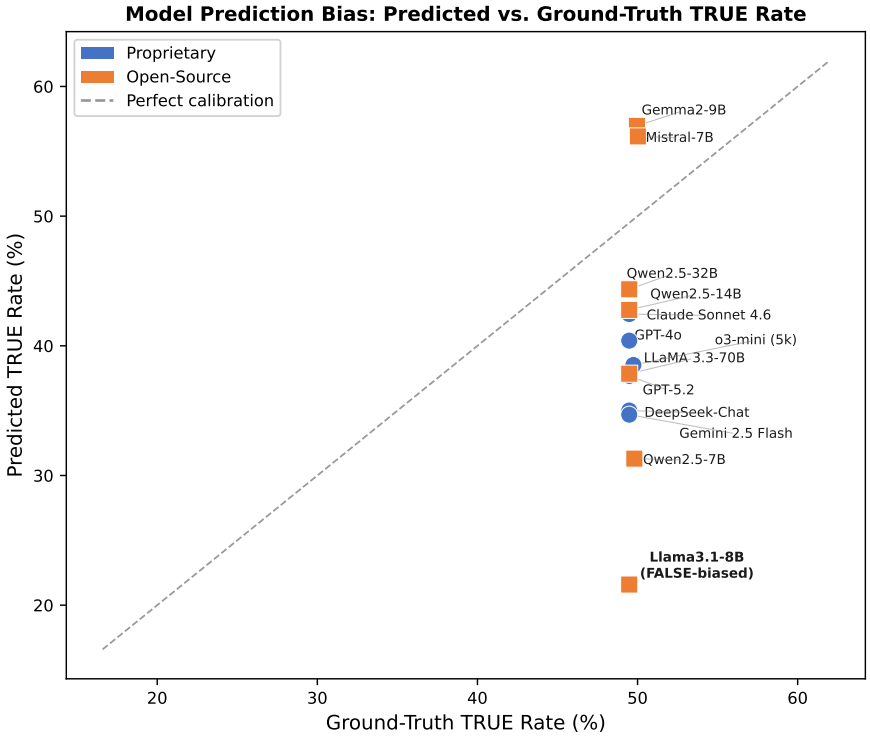


Figure 4: Predicted vs. ground-truth TRUE rate (49.5%). LLaMA 3.1-8B predicts TRUE only 21.6%.

learn from text (logical rule-following) and what requires numerical grounding (parameter-dependent dynamics).

This complements findings from PrOntoQA (Saparov & He, 2023), which showed that LLMs struggle with longer synthetic reasoning chains. Our results show that even short chains succeed when premises are given (FOL inference), but the difficulty shifts from chain length to premise availability: models cannot generate the quantitative facts needed for bifurcation reasoning.

Consistency failures expose fragile retrieval. Consistency_paraphrase (MCC = 0.25) and perturbation (MCC = 0.26) are not knowledge gaps. Models answer "Is Lorenz-63 chaotic?" correctly at the atomic level (MCC = 0.32–0.62) but flip when the same fact is rephrased. The knowledge exists; the retrieval is sensitive to surface form.

Proprietary vs. open-source: not a monolithic gap. The per-family decomposition (Figure 3) shows the gap concentrates on cross-indicator reasoning (+0.40) and consistency (+0.22), while formal deduction and perturbation robustness show near-parity. Qwen 2.5-32B's MCC = 0.91 on indicator_diagnostic (interpreting Lyapunov exponents, permutation entropy, MEGNO) exceeds every proprietary model, suggesting that training data composition matters more than the proprietary/open-source divide for quantitative threshold reasoning.

MaxSAT axiom repair. We apply a MaxSAT post-processor (RC2 (Ye et al., 2023)) that repairs per-system predicate assignments to satisfy all 78 axiom edges with minimal prediction flips. On atomic questions, repair eliminates all FOL violations and improves MCC substantially for weak models: LLaMA 3.1-8B gains +0.11 MCC (0.22→0.33) with 11.2% of predictions flipped; Mistral-7B gains +0.11 (0.17→0.28) with 19.0% flipped. Claude Sonnet 4.6 is barely affected (−0.006 MCC). When we propagate the repaired predicates to com-

positional families (multi_hop, fol_inference), the picture reverses: repair degrades MCC on these families (e.g., −0.20 on multi_hop for LLaMA), because compositional questions encode reasoning that goes beyond per-predicate consistency. This reveals a separation between two error types: axiom-inconsistent errors (fixable by constraint enforcement) and reasoning errors (requiring deeper inference). Solver augmentation helps the first type but not the second, identifying a precise boundary for hybrid LLM-solver approaches.

Chain-of-thought preliminary. We tested CoT prompting on the two hardest families (regime_transition, cross_indicator; 135 questions) using LLaMA 3.1-8B. CoT dramatically increased the invalid rate: 66/68 regime_transition responses and 57/67 cross_indicator responses were unparseable (the model generates reasoning text but fails to produce a final TRUE/FALSE). Among the few parseable CoT responses, regime_transition remained at MCC = 0.0. This suggests that for small models, CoT on hard scientific reasoning families introduces a format-compliance problem without solving the underlying reasoning gap.

### 6.1 Limitations

Three families have $N < 100$ (regime_transition, cross_indicator, extended_systems); results on these carry high variance. The CoT experiment uses only one small model (8B); larger models with CoT may behave differently. The TRUE/FALSE format cannot distinguish correct reasoning from correct guessing. Well-known systems (Lorenz-63) may be memorized from pretraining rather than reasoned about; the extended_systems ceiling effect is consistent with this.

## 7 Conclusion

ChaosBench-Logic v2 and the CARE evaluation protocol together reveal that apparent LLM reasoning performance hides three pathologies: prior collapse (LLaMA 3.1-8B achieves 60% accuracy with only 32% TPR), surface-form fragility (consistency MCC = 0.25), and inability to reason about parameter-dependent dynamics (regime transition MCC = 0.05). The knowledge-type boundary between rule-following and parametric reasoning does not close with scale. Our MaxSAT repair experiment reveals two distinct error types: axiom-inconsistent errors (fixable by constraint enforcement, +0.11 MCC on atomic questions for weak models) and reasoning errors on compositional families (degraded by repair), identifying a precise boundary for solver-augmented approaches. Three directions follow. First, chain-of-thought evaluation: preliminary results from v1 of this benchmark (Thomas, 2026) found that CoT decreased overall accuracy by 2–6 percentage points for both GPT-4 and LLaMA-3, suggesting that explicit reasoning introduces errors on scientific questions where zero-shot retrieval is more reliable. Whether this pattern holds on v2's harder families (regime_transition, cross_indicator) remains an open question. Second, deeper solver integration: our MaxSAT repair fixes axiom-inconsistent errors but not reasoning errors; coupling LLMs with numerical integrators could address the compositional gap. Third, fine-tuning on scientific corpora could test whether consistency failures reflect missing knowledge or architectural limitations. The benchmark, CARE protocol, and released artifacts are publicly available at https://github.com/11NOel11/ChaosBench-Logic and https://huggingface.co/datasets/11NOel11/ChaosBench-Logic.

## Reproducibility Statement

All evaluations use deterministic inference (temperature = 0 where supported; reasoning models are deterministic by design). Metrics are computed from raw confusion matrices in prediction logs; every number was verified against these logs. The dataset is generated deterministically from the FOL axiom system. Code, canonical dataset files, and released artifacts are available at https://github.com/11NOel11/ChaosBench-Logic and https://huggingface.co/datasets/11NOel11/ChaosBench-Logic.

## References

Christoph Bandt and Bernd Pompe. Permutation entropy: A natural complexity measure for time series. Physical Review Letters, 88(17):174102, 2002.

Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1):6, 2020.

Pablo M Cincotta and Carles Simó. Simple tools to study global dynamics in non-axisymmetric galactic potentials – I. Astronomy and Astrophysics Supplement Series, 147:205–228, 2000.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.

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

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In Advances in Neural Information Processing Systems (Datasets and Benchmarks Track), 2021.

Georg A Gottwald and Ian Melbourne. On the implementation of the 0-1 test for chaos. SIAM Journal on Applied Dynamical Systems, 8(1):129–145, 2009.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Schmitt, Hinrich Schütze, Volker Tresp, and Nanyun Peng. FOLIO: Natural language reasoning with first-order logic. arXiv preprint arXiv:2209.00840, 2022.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. Advances in Neural Information Processing Systems, 34:37914–37927, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 3214–3252, 2022.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 3622–3628, 2020.

Edward N Lorenz. Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20 (2):130–141, 1963.

Brian W Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) – Protein Structure, 405(2): 442–451, 1975.

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

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research, 2023.

Steven H Strogatz. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. CRC Press, 2nd edition, 2018.

Noel Thomas. Chaosbench-logic: A benchmark for logical and symbolic reasoning on chaotic dynamical systems. 2026.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv preprint arXiv:2307.10635, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations, 2023b.

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

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. SatLM: Satisfiability-aided language models using declarative prompting. In Advances in Neural Information Processing Systems, volume 36, 2023.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A reading comprehension dataset requiring logical reasoning. In International Conference on Learning Representations, 2020.

# A Full Leaderboard and Subset Evaluations

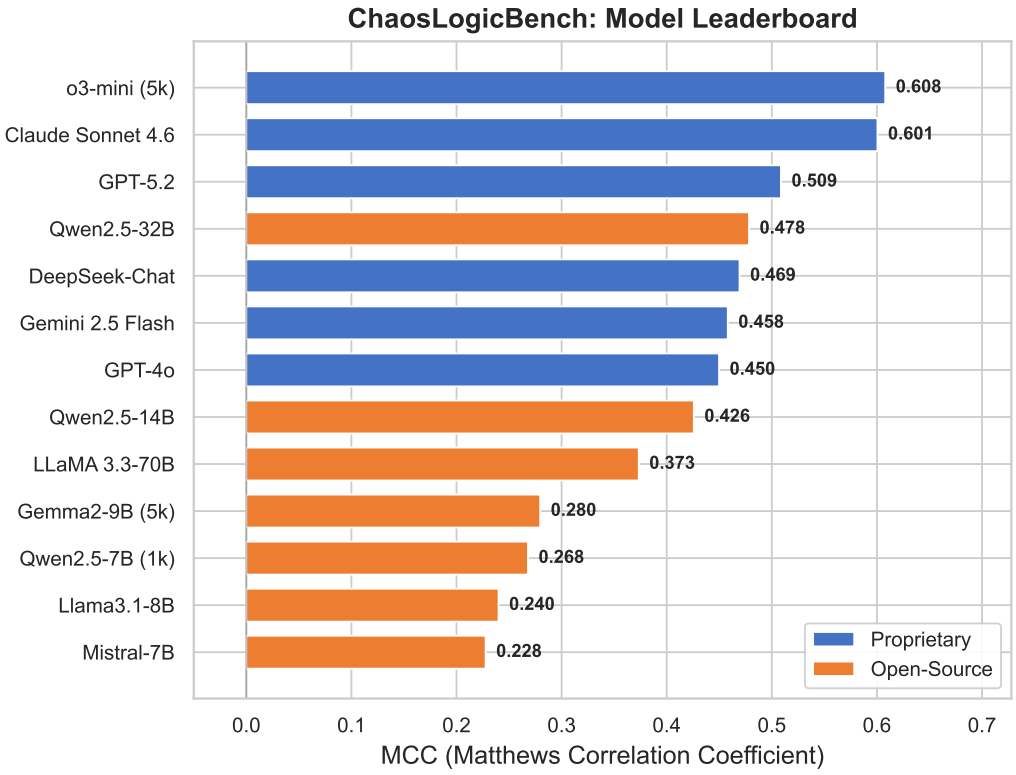


Figure 5: MCC by model. Blue: proprietary. Orange: open-source.

Table 4: Subset evaluations. 5k tracks full-canonical MCC within 0.01 (validated on 3 models with both); 1k has higher variance.

| Model | Type | Bal. Acc | MCC | N |
|---|---|---|---|---|
| o3-mini | Prop. | 0.796 | 0.608 | 5,000 |
| Gemma2-9B | OSS | 0.639 | 0.280 | 5,000 |
| GPT-4o-mini | Prop. | 0.623 | 0.272 | 1,000 |
| Qwen 2.5-7B | OSS | 0.624 | 0.268 | 1,000 |

# B Model Size Scaling

# C Subset Validation

For models with both evaluations (Qwen 2.5-32B, Qwen 2.5-14B, Mistral-7B), mean $|\Delta\text{MCC}| < 0.01$ overall; per-family mean $|\Delta\text{MCC}| = 0.050$ for families with $N \geq 30$ in the 5k subset.

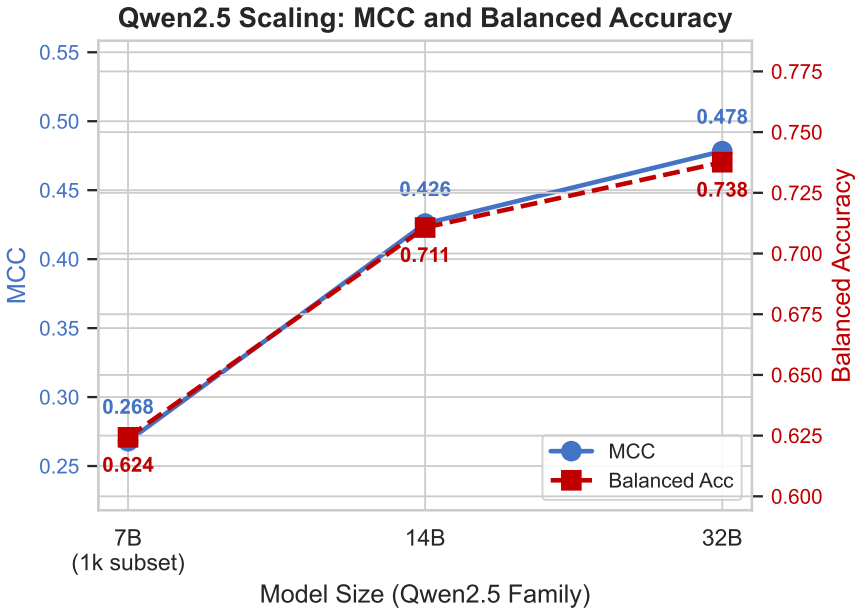


Figure 6: Scaling within Qwen 2.5. MCC increases from 0.27 (7B, 1k subset) to 0.48 (32B), but regime_transition remains near-random at all scales.

Table 5: Example questions with ground truth (GT) and predictions. Son. = Claude Sonnet 4.6; Qw. = Qwen 2.5-32B; Ll. = LLaMA 3.3-70B. Incorrect predictions in bold.

| Family | Question (abbreviated) | GT | Son. | Qw. | Ll. |
|---|---|---|---|---|---|
| multi_hop | "FluidTrampoline is strongly mixing. Strongly mixing ⇒ weakly mixing ⇒ ergodic ⇒ bounded. Is it bounded?" | T | T | F | F |
| regime | "At $\alpha$=15.6, is Chua's circuit chaotic?" | T | T | T | T |
| fol_inf | "Given Torus is stat. predictable, must it be bounded?" | T | T | T | F |
| adv_misl | "Given bounded, can Rössler be weak mixing?" | T | T | F | F |

## D Family Discrimination

## E Question Examples

## F Axiom Specification

The six primary regime axioms:

1. Chaotic ⇒ Deterministic ∧ PosLyap ∧ Sensitive ∧ PointUnpredictable ∧ StatPredictable ∧ Mixing; excludes Random, Periodic, QuasiPeriodic, FixedPointAttr.
2. Random excludes Deterministic, Chaotic, QuasiPeriodic, Periodic.
3. QuasiPeriodic ⇒ Deterministic ∧ Bounded; excludes Chaotic, Random, Periodic, FixedPointAttr.
4. Periodic ⇒ Deterministic ∧ Bounded; excludes Chaotic, Random, QuasiPeriodic, StrangeAttr.
5. FixedPointAttr ⇒ Deterministic; excludes Chaotic, Random, QuasiPeriodic, Periodic, StrangeAttr.
6. Deterministic excludes Random.

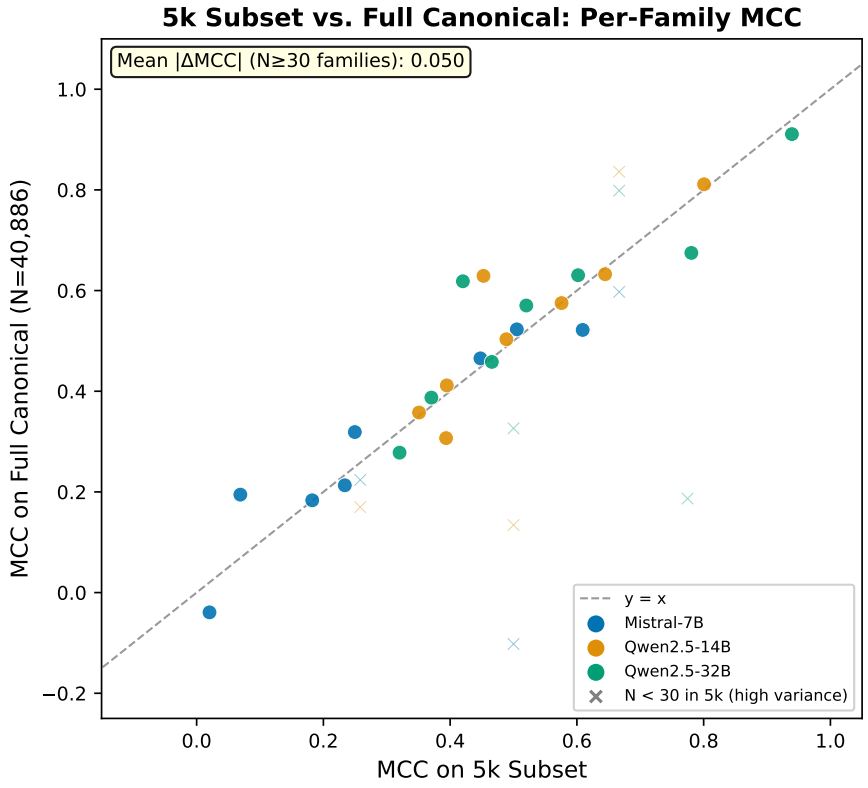


Figure 7: 5k vs. full canonical per-family MCC. Circles: $N \geq 30$. Crosses: $N < 30$ (high variance).

Additional edges: PosLyap ⇒ Sensitive ⇒ PointUnpredictable; StrangeAttr ⇒ Dissipative ∧ Bounded; Mixing ⇒ Ergodic ⇒ Bounded; HyperChaotic ⇒ Chaotic ∧ StrangeAttr ∧ Dissipative; Conservative ⇒ Bounded ∧ Ergodic; StrongMixing ⇒ WeakMixing ⇒ Ergodic; ContinuousTime ↔ ¬DiscreteTime; Forced ↔ ¬Autonomous; DelaySystem ⇒ ContinuousTime.

## G Confusion Matrices

Table 6: Regime_transition (N = 68) confusion matrices.

| Model | TP | FP | TN | FN | MCC | Bal. Acc |
|---|---|---|---|---|---|---|
| LLaMA 3.3-70B | 9 | 17 | 20 | 22 | −0.173 | 0.415 |
| Mistral-7B | (from aggregate) | | | | −0.102 | – |
| Claude Sonnet 4.6 | 15 | 5 | 32 | 16 | +0.381 | 0.674 |

## H Invalid Rates

Most models produce zero invalids. Mistral-7B has the highest rate at 1.1%; LLaMA 3.1-8B <0.01%; all others 0.0%.

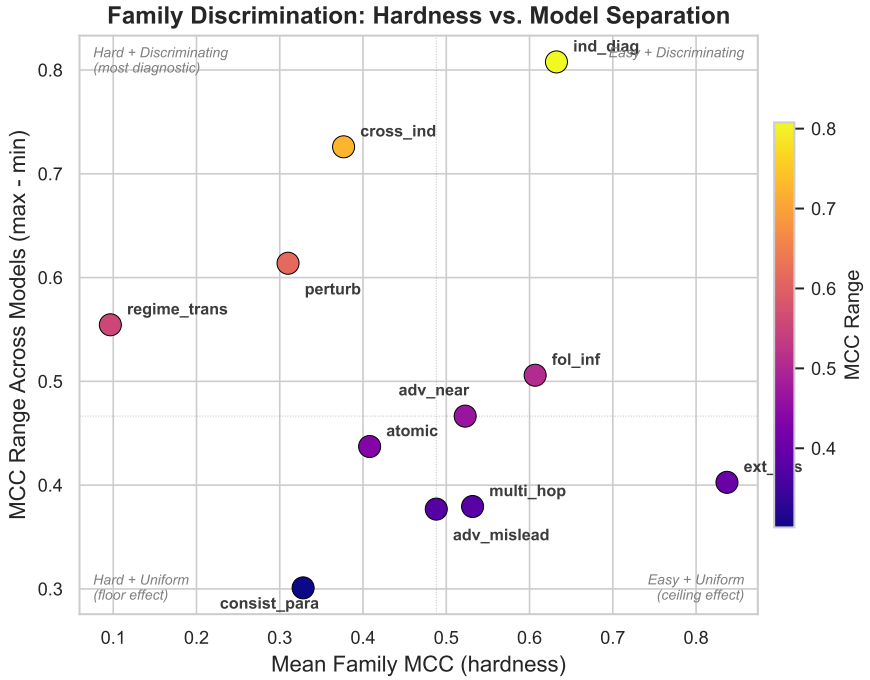


Figure 8: Mean MCC vs. MCC range across 10 models. Upper-left: hard and discriminating. Lower-left: floor effects.

# I Prompt Template

> Answer the following question about the dynamical system. Reply with only TRUE or FALSE.
> Question: [question text]
> Answer:

No system prompt, few-shot examples, or CoT instructions. Temperature = 0 for all models that accept the parameter. Reasoning models (o3-mini, GPT-5.2) do not accept a temperature argument; their outputs are internally deterministic via the reasoning process. Max tokens: 16 for most models; 1024 for o3-mini and GPT-5.2 (reasoning token budget), and Gemini 2.5 Flash (thinking process).