# ChaosBench-Logic: A Benchmark for Logical and Symbolic Reasoning on Chaotic Dynamical Systems

**Noel Thomas**

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
noel.thomas@mbzuai.ac.ae

## Abstract

Large language models (LLMs) excel at natural language tasks but remain brittle in domains requiring precise logical and symbolic reasoning. Chaotic dynamical systems provide an especially demanding test because chaos is deterministic yet frequently conflated with randomness, complexity, or mere nonlinearity. We introduce CHAOSBENCH-LOGIC, a benchmark that evaluates LLM reasoning across 30 diverse dynamical systems using a unified first-order logic ontology. Each system is annotated with truth assignments for 11 semantic predicates, and 621 questions are generated across seven reasoning categories, including multi-hop implications, cross-system analogies, counterfactual reasoning, bias probes, and multi-turn dialogues. We define metrics for logical accuracy, implication consistency, dialogue coherence, and contradiction, and we release an open-source evaluation pipeline. Initial experiments show that frontier LLMs such as GPT-4, Claude 3.5 Sonnet, Gemini 2.5 Flash, and the open-source LLaMA-3 70B achieve 91–94% per-item accuracy, yet still score 0% on compositional items and exhibit fragile global coherence: dialogue-level accuracy ranges from 53.1% (GPT-4 CoT) to 75.5% (LLaMA-3 zeroshot). CHAOSBENCH-LOGICprovides a rigorous testbed for diagnosing these failures and a foundation for developing neuro-symbolic approaches that improve scientific reasoning in LLMs.

**Code (GitHub)** — https://github.com/11NOel11/ChaosBench-Logic

**Dataset (Hugging Face)** — https://huggingface.co/datasets/11NOel11/ChaosBench-Logic

## 1 Introduction

Large language models have transformed natural language processing, demonstrating remarkable abilities in translation, summarisation, question answering, and even code synthesis. Despite this progress, they remain surprisingly brittle when confronted with tasks requiring precise logical or symbolic reasoning. In scientific domains such as dynamical systems and chaos theory, accurate reasoning is vital: one must distinguish whether a system is chaotic, periodic, quasi-periodic, or random; infer implications of qualitative properties; and reason about parameter perturbations and bifurcations. Errors in such reasoning can lead to incorrect scientific conclusions and undermine trust in LLM-based scientific tools.

Chaotic dynamics provide a stringent test of reasoning because "chaos" is often misunderstood even by experts. Chaos is deterministic yet exhibits sensitive dependence on initial conditions, positive Lyapunov exponents, and complex attractor geometry [19, 14, 3]. It is *not* equivalent to randomness, nonlinearity, or complexity; confusing these concepts causes systematic misinterpretations of physical systems. Existing LLM benchmarks do not adequately test this mix of domain semantics and formal reasoning depth. Our goal is therefore to develop a benchmark that systematically evaluates and analyses LLM reasoning about chaotic and non-chaotic systems under formal logical constraints.

### 1.1 Contributions

Our work makes the following key contributions:

1. **CHAOSBENCH-LOGIC Benchmark.** A benchmark of 621 reasoning questions over 30 dynamical systems spanning ODEs, maps, PDEs, neuronal and chemical models, and stochastic processes [20, 21].
2. **first-order logic Ontology.** A unified logical vocabulary (11 predicates) and a compact axiom system capturing widely accepted implications between chaos, determinism, Lyapunov behavior, randomness, attractor types, and predictability.
3. **Seven Task Families.** Atomic QA, multi-hop implications, cross-system analogies and non-analogies, counterfactual reasoning, bias probes, multi-turn dialogues, and hard compositional questions.
4. **Metrics and Pipeline.** Reproducible evaluation for accuracy, implication consistency, dialogue coherence, contradiction detection, and axiom-violation counts.
5. **Empirical Diagnosis.** Frontier and open-source LLMs achieve high per-item accuracy (91–94%) but fail on compositional items (0%) and show fragile multi-turn coherence.

## 2 Related Work

**Logical and Symbolic Reasoning in LLMs.** Recent work assesses LLMs on deductive and symbolic tasks, includ-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ing BIG-Bench [18], ReClor [25], LogiQA [10], and related logic benchmarks. These datasets reveal persistent brittleness in multi-step inference, contradiction avoidance, entailment, and reasoning under explicit constraints. Several efforts introduce structured prompting (e.g., chain-of-thought) or tool-augmented reasoning, yet models still hallucinate intermediate steps or fail to maintain coherence across turns [24]. CHAOSBENCH-LOGICextends this line of inquiry to a scientific domain where reasoning must stay consistent with formal definitions of chaos-related semantics.

**Scientific QA and Domain Reasoning.** Benchmarks such as SciBench [23] and MATH [6] evaluate quantitative or symbolic scientific problem solving, highlighting limitations in equation manipulation and multi-step reasoning. However, none target dynamical systems theory, nor do they ground questions in a shared logical ontology. CHAOSBENCH-LOGICdiffers by anchoring all questions in a unified first-order logic structure and scoring models for *logical coherence* rather than only final-answer plausibility.

**Chaos and Machine Learning.** Prior ML-for-chaos work primarily evaluates forecasting or attractor reconstruction (e.g., reservoir computing and deep learning baselines) rather than formal reasoning about qualitative regimes [15, 4]. CHAOSBENCH-LOGICcomplements these benchmarks by focusing on symbolic and logical competence: distinguishing chaos from randomness and enforcing consistent implication structure across diverse systems.

**Bias and Misconception Diagnostics.** Evaluations such as TruthfulQA [9] examine systematic model errors and truthfulness failures. Yet domain-specific misconceptions (e.g., chaos–randomness confusion, nonlinearity–chaos conflation) remain largely unexplored. CHAOSBENCH-LOGICintroduces explicit bias-probe families tailored to chaos theory semantics.

# 3 Methods

This section formalizes CHAOSBENCH-LOGIC, describes the ontology and axiom system, details dataset construction, and defines evaluation metrics.

## 3.1 Problem Definition

We consider reasoning over a collection of dynamical systems $\mathcal{S} = \{s_1, \ldots, s_{30}\}$ including continuous-time flows, discrete-time maps, PDEs, neuronal oscillators, chemical reaction models, and stochastic processes.

Each system $s \in \mathcal{S}$ is associated with semantic properties expressed as unary logical predicates from a fixed vocabulary $\mathcal{P}$. Given a natural-language query about a system $s$, the model must predict the logically correct answer according to the system's ground-truth property assignment and a global axiom system.

## 3.2 Logical Predicate Vocabulary

CHAOSBENCH-LOGICdefines 11 unary predicates, each mapping a system $s$ to a boolean:

- $\mathsf{Chaotic}(s)$
- $\mathsf{Deterministic}(s)$
- $\mathsf{PosLyap}(s)$ (positive largest Lyapunov exponent)
- $\mathsf{Sensitive}(s)$
- $\mathsf{StrangeAttr}(s)$
- $\mathsf{PointUnpredictable}(s)$
- $\mathsf{StatPredictable}(s)$
- $\mathsf{QuasiPeriodic}(s)$
- $\mathsf{Random}(s)$ (true randomness / stochasticity)
- $\mathsf{FixedPointAttr}(s)$
- $\mathsf{Periodic}(s)$

## 3.3 Predicate Semantics and Annotation Protocol

To make the benchmark auditable and to reduce ambiguity, we treat each predicate as a *semantic label* over a named system (or canonical regime) rather than as an invitation for models to "interpret" definitions ad hoc. In particular, the benchmark is designed so that a system can appear random in its trajectories and still be labeled deterministic, reflecting the core distinction between *stochasticity* and *sensitive determinism*.

We use the following annotation principles:

- **Determinism vs. randomness.** $\mathsf{Deterministic}(s)$ indicates that the system evolution is fully specified by its state and update rule. $\mathsf{Random}(s)$ is reserved for systems with inherent stochasticity (e.g., explicit noise terms or stochastic differential equations).
- **Chaos as a regime label.** $\mathsf{Chaotic}(s)$ indicates a canonical parameter regime widely treated as chaotic in the literature for that named system. We avoid borderline regimes and remove disputed cases during curation.
- **Predictability predicates.** $\mathsf{PointUnpredictable}(s)$ captures practical loss of long-horizon point forecasts even under perfect modeling, while $\mathsf{StatPredictable}(s)$ reflects that coarse statistical structure can remain predictable or stable over time (e.g., invariant measures).
- **Attractor-type predicates.** $\mathsf{Periodic}(s)$, $\mathsf{QuasiPeriodic}(s)$, and $\mathsf{FixedPointAttr}(s)$ are mutually exclusive *indicators* of common non-chaotic behavior classes, used to generate non-analogies and "do not overgeneralize" items.

These principles make CHAOSBENCH-LOGICcloser to a controlled reasoning benchmark than a definition-discovery task: models are evaluated on logical consistency under a fixed ontology and ground truth.

## 3.4 Global first-order logic Axiom System

CHAOSBENCH-LOGICspecifies a set $\Phi$ of global axioms encoding widely used implications in dynamical systems [19, 14, 3]. We write each implication explicitly:

**Chaos axioms.**

$$\forall s: \ \mathsf{Chaotic}(s) \Rightarrow \mathsf{Deterministic}(s), \tag{1}$$
$$\forall s: \ \mathsf{Chaotic}(s) \Rightarrow \mathsf{PosLyap}(s), \tag{2}$$
$$\forall s: \ \mathsf{Chaotic}(s) \Rightarrow \mathsf{Sensitive}(s), \tag{3}$$
$$\forall s: \ \mathsf{Chaotic}(s) \Rightarrow \mathsf{PointUnpredictable}(s), \tag{4}$$
$$\forall s: \ \mathsf{Chaotic}(s) \Rightarrow \mathsf{StatPredictable}(s), \tag{5}$$
$$\forall s: \ \mathsf{Chaotic}(s) \Rightarrow \neg\mathsf{Random}(s). \tag{6}$$

**Lyapunov & predictability.**

$$\forall s: \ \mathsf{PosLyap}(s) \Rightarrow \mathsf{Sensitive}(s), \tag{7}$$
$$\forall s: \ \mathsf{PosLyap}(s) \Rightarrow \mathsf{PointUnpredictable}(s). \tag{8}$$

**Attractor-type implications.**

$$\forall s: \ \mathsf{StrangeAttr}(s) \Rightarrow \mathsf{Chaotic}(s), \tag{9}$$
$$\forall s: \ \mathsf{Periodic}(s) \Rightarrow \neg\mathsf{Chaotic}(s), \tag{10}$$
$$\forall s: \ \mathsf{QuasiPeriodic}(s) \Rightarrow \neg\mathsf{Chaotic}(s), \tag{11}$$
$$\forall s: \ \mathsf{FixedPointAttr}(s) \Rightarrow \neg\mathsf{Chaotic}(s). \tag{12}$$

**Randomness vs determinism.**

$$\forall s: \ \mathsf{Random}(s) \Rightarrow \neg\mathsf{Deterministic}(s). \tag{13}$$

**Design choice: no converse rules.** Crucially, we *do not* assume reverse implications (e.g., Sensitive $\Rightarrow$ Chaotic). This avoids overspecification and forces models to respect directionality: many scientific misconceptions arise from treating one-way implications as equivalences.

### 3.5 Ground-Truth Closure and Answer Derivation

The benchmark ground truth for each system is a truth assignment for the 11 predicates that is consistent with $\Phi$. For implication questions, we compute the correct answer by applying *forward chaining* under $\Phi$.

Let $\mathcal{A}_s$ denote the set of all literals known true/false for a system $s$. We define the logical closure under $\Phi$ as:

$$\mathrm{Cl}_\Phi(\mathcal{A}_s) = \mathrm{lfp}\Big(\mathcal{A}_s \cup \{\, \ell \ : \ (\ell_1 \wedge \cdots \wedge \ell_k \Rightarrow \ell) \in \Phi, \ell_1, \ldots, \ell_k \in \mathcal{A}_s\}\Big), \tag{14}$$

where lfp is the least fixed point reached by repeatedly applying all applicable implications. Because the system annotations satisfy $\Phi$, the closure remains consistent. For model outputs, however, closure and axiom checks expose contradictions and implication violations.

### 3.6 System Coverage

CHAOSBENCH-LOGICincludes 30 systems across five broad classes. Table 1 provides an overview.

We include canonical systems spanning flows and maps (e.g., Lorenz and Rössler systems, Hénon and logistic maps) as well as PDE and stochastic processes [11, 16, 7, 12, 17, 8, 22].

Table 1: Overview of dynamical systems included in CHAOSBENCH-LOGIC. Each system is annotated with 11 semantic predicates and participates in first-order logic reasoning.

| Class | Examples | Chaos? |
|---|---|---|
| Chaotic ODEs | Lorenz-63, Lorenz-84, Lorenz-96, Rössler, Chen, HR neuron | yes |
| Non-chaotic ODEs | Simple harmonic oscillator (SHM), Van der Pol, Brusselator, FitzHugh–Nagumo | no |
| Maps | Logistic, Hénon, Ikeda, Baker, Arnold Cat, Standard Map, Circle Map | mixed |
| PDEs | Kuramoto–Sivashinsky (chaotic), Sine–Gordon (integrable) | mixed |
| Stochastic | Ornstein–Uhlenbeck | random, non-chaotic |

### 3.7 Dataset Construction Pipeline

The benchmark is produced via a four-stage pipeline:

**(1) System encoding.** Each system is encoded in a structured JSON file containing equations or update rules, canonical parameter regimes, short natural-language summaries, and predicate truth assignments. This representation supports both human inspection and programmatic question generation.

**(2) Template-based generation with semantic constraints.** We design seven task families with 70+ templates. Templates are not free-form: each includes explicit semantic constraints (e.g., "only apply this template to systems with $\mathsf{Random}(s) = \mathsf{TRUE}$") to avoid ill-posed questions. We also vary lexical surface forms and question structure to reduce shallow cue exploitation.

**(3) Human curation for correctness and ambiguity removal.** All candidates are manually checked for logical correctness under $\mathrm{Cl}_\Phi(\cdot)$, linguistic clarity, and consistency with the intended misconception being tested (for bias probes). Ambiguous items (e.g., parameter regimes with disputed behavior) are removed.

**(4) Final filtering and consistency checks.** We automatically filter questions that violate applicability rules or create logical self-conflicts, then perform a final manual pass. In addition to basic sanity checks, we verify that each item is answerable from the system labels and $\Phi$ alone (i.e., it does not require numerical simulation).

### 3.8 Question Format and Metadata

Each benchmark item includes a system identifier and a natural-language question, plus structured metadata used in evaluation. The current release includes fields such as `system_id`, `family`, `template_id`, and (for dialogues) `dialogue_id` and `turn_index`. This makes it possible to compute aggregate metrics per family and to audit failures at the level of a specific template or system.

For compositional items, the metadata explicitly records which sub-skills are required (e.g., implication chaining plus analogy rejection), enabling more fine-grained analysis than a single global accuracy score.

### 3.9 Reasoning Task Families

CHAOSBENCH-LOGICcovers seven reasoning types:

1. **Atomic Logical QA:** single-step predicate queries.
2. **Multi-Hop Implication Reasoning:** forward-chaining under $\Phi$.
3. **Cross-System Analogical / Non-Analogical Reasoning:** transfer and rejection.
4. **Counterfactual Parameter Reasoning:** bifurcations and perturbations.
5. **Bias Probes:** adversarial misconception tests.
6. **Advanced Multi-Turn Dialogues:** cross-turn coherence and contradiction avoidance.
7. **Hard Concept-Synthesis / Compositional:** multiple reasoning modes combined.

Table 2 summarizes what each family measures.

### 3.10 Multi-Turn Dialogue Design

Each dialogue is grouped under a `dialogue_id`, with 3–6 turns. Correctness requires per-turn accuracy *and* intra-dialogue coherence: answers must not contradict earlier commitments under $\Phi$. This targets persistent reasoning and belief-state stability.

A key design choice is that dialogue turns intentionally mix "easy" atomic questions with implication-driven follow-ups. This stresses whether a model can (i) answer a local question, (ii) remember its own earlier commitment, and (iii) apply the same commitment consistently when queried from a different angle later in the conversation.

Table 2: What each task family tests and why it is non-trivial for LLMs.

| Family | What it measures (typical failure) |
|---|---|
| Atomic QA | Correctly recall system-level truth values (confident hallucination). |
| Multi-hop | Apply directionality of implications (converse fallacy). |
| Analogy | Transfer only the shared structure (surface-form analogy bias). |
| Counterfactual | Reason about regime change vs invariance (overgeneralize chaos). |
| Bias probes | Resist misleading cues ("chaos = random", "PDE = chaotic"). |
| Dialogues | Maintain consistent belief state over turns (belief drift). |
| Compositional | Integrate multiple constraints simultaneously (local-but-not-global reasoning). |

### 3.11 Bias Families

We identify six systematic reasoning biases:

1. Chaos–Randomness Bias
2. Chaos–Nonlinearity Bias
3. Complexity–Chaos Bias
4. Fractal–Chaos Bias
5. PDE–Chaos Overgeneralisation
6. Name/Analogy Bias ("Lorenz-like $\rightarrow$ chaotic")

These biases are encoded at the template level in the current release and can be refined into per-item labels in future versions.

### 3.12 Evaluation Metrics

We introduce metrics that separate *local* correctness from *global* coherence.

**Logical accuracy.** Given $N_{\text{eval}}$ questions with valid normalized predictions:

$$\mathrm{Acc} = \frac{1}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} \mathrm{I}\big(\hat{y}_i = y_i^*\big). \tag{15}$$

**Dialogue accuracy.** For a dialogue $d$ with turns $t = 1, \ldots, T_d$:

$$\mathrm{DialogueAcc}(d) = \mathrm{I}\big(\forall t,\ \hat{y}_{d,t} = y_{d,t}^*\big). \tag{16}$$

This metric is strict by design: a single incoherent turn collapses a whole dialogue to 0.

**Contradiction & axiom-violation checks.** Beyond correctness, we evaluate whether a model's *set of commitments* is internally consistent. We flag contradictions (e.g., predicting both $\mathsf{Chaotic}(s)$ and $\mathsf{Random}(s)$) and count violations of implications in $\Phi$ within a dialogue. These diagnostics explain *how* and *why* dialogues fail even when per-turn answers often look locally plausible.

Table 3: Overall performance on CHAOSBENCH-LOGIC. Acc is per-item accuracy; DlgAcc is the fraction of dialogues where all turns are correct.

| Model | Mode | Acc (%) | DlgAcc (%) |
|---|---|---|---|
| GPT-4 | Zeroshot | 94.0 | 69.4 |
| GPT-4 | CoT | 88.2 | 53.1 |
| Claude 3.5 Sonnet | Zeroshot | 91.6 | 67.3 |
| Gemini 2.5 Flash | Zeroshot | 91.9 | 71.4 |
| LLaMA-3 70B | Zeroshot | 91.6 | 75.5 |
| LLaMA-3 70B | CoT | 89.5 | 65.3 |

# 4 Experimental Setup and Baselines

**Models.** We evaluate representative frontier and open-source LLMs: GPT-4 (OpenAI), Claude 3.5 Sonnet (Anthropic), Gemini 2.5 Flash (Google), and LLaMA-3 70B (Meta) [13, 2, 5, 1]. All models are queried through their official APIs or hosted endpoints.

**Prompting.** Zero-shot prompts request a final YES/NO decision. For chain-of-thought (CoT) prompting, we ask the model to write reasoning and end with `FINAL_ANSWER: YES/NO` [24]. The evaluation script extracts and normalizes the final label.

**Hyperparameters and determinism.** All models are run at temperature 0 (or nearest deterministic setting) to reduce sampling variance and emphasize reasoning consistency failures.

**Evaluation pipeline.** Responses are normalized using a robust pattern matcher. If a response contains multiple answers, we prefer the final explicit marker (CoT) or the last YES/NO token (zero-shot). This avoids over-crediting models that hedge or self-correct mid-output.

# 5 Results

We now present empirical results on CHAOSBENCH-LOGIC. All accuracies are computed over items where a valid normalized answer could be extracted; in our final runs, this is 621/621 items for all configurations.

## 5.1 Overall Logical Performance

Table 3 summarizes performance. All evaluated models achieve 91–94% per-item accuracy, yet dialogue-level coherence is substantially lower.

**Why DlgAcc is much lower than Acc.** Dialogue accuracy compounds errors: for a dialogue of length $T$, even an independent per-turn accuracy $p$ yields an "all-correct" probability of approximately $p^T$. With $p \approx 0.92$ and $T \in [3, 6]$, this ranges from $0.92^3 \approx 0.78$ to $0.92^6 \approx 0.61$, matching the observed DlgAcc range (53–76%). Crucially, *real* dialogue errors are not independent: models often drift in their interpretation of a system across turns, causing clustered failures rather than isolated mistakes.

Table 4: Task-specific best models on CHAOSBENCH-LOGIC. Accuracies are over items with a valid normalised answer.

| Task family | Best model | Acc (%) |
|---|---|---|
| Atomic logical QA | GPT-4 (Zeroshot) | 96.1 |
| Bias probes | GPT-4 (Zeroshot) | 93.0 |
| Counterfactuals | GPT-4 (Zeroshot) | 97.4 |
| Multi-hop reasoning | GPT-4 (Zeroshot) | 97.1 |
| Multi-turn (per-turn) | LLaMA-3 (Zeroshot) | 93.9 |
| Dialogue (dialogue-level) | LLaMA-3 (Zeroshot) | 75.5 |
| Compositional | All models | 0.0 |

**Local competence vs. global coherence.** The high Acc values show that most *single* questions are solvable for current LLMs. The lower DlgAcc values indicate that models struggle to maintain a stable belief state across turns—even when the required beliefs are simple boolean properties with deterministic implications.

## 5.2 Task-Family Highlights and the "0% Compositional" Gap

Table 4 reports the best-performing model for key task families. Models are strong on many atomic and multi-hop items but fail completely on compositional questions.

**What makes compositional items qualitatively harder.** Compositional questions deliberately combine multiple constraints, such as: (i) chaining two or more implications under $\Phi$; (ii) rejecting a misleading analogy cue; (iii) applying a counterfactual change shown elsewhere in the dialogue; and (iv) producing an answer consistent with prior turns. In practice, models often succeed on each component *in isolation* but fail when the components must be integrated into a single coherent decision.

**A key observed failure mode: converse assumptions.** Because $\Phi$ contains only one-way implications, compositional items often punish the common LLM heuristic "if A implies B, then B implies A." For example, models may incorrectly infer $\mathsf{Chaotic}(s)$ from $\mathsf{Sensitive}(s)$ or from $\mathsf{PosLyap}(s)$ even though our ontology does not assume those converses. This is not a trick: it reflects real scientific reasoning discipline about implication directionality.

## 5.3 Consistency Diagnostics Beyond Accuracy

A major advantage of CHAOSBENCH-LOGICis that it makes inconsistency measurable. Two models can achieve similar per-item accuracy while differing sharply in *how* they fail. In particular, we observe that some errors are "local" (a single wrong predicate assignment) while others represent *global* constraint failures (e.g., asserting both $\mathsf{Random}(s)$ and $\mathsf{Deterministic}(s)$, or implying chaos from a non-converse symptom).

These failures matter operationally: in downstream scientific workflows, a locally incorrect answer may be caught by a user, but a globally inconsistent belief state can produce confident yet mutually incompatible explanations across

Table 5: CoT impact on overall and dialogue-level performance.

| Model | ΔAcc (pp) | ΔDlgAcc (pp) | Trend |
|---|---|---|---|
| GPT-4 | −5.8 | −16.3 | worse |
| LLaMA-3 70B | −2.1 | −10.2 | worse |

time. This is precisely the failure mode that multi-turn dialogues and axiom-violation tracking are designed to surface.

### 5.4 Effect of Chain-of-Thought Prompting

CoT is frequently assumed to improve reasoning [24], but on CHAOSBENCH-LOGICit reduces both accuracy and dialogue-level coherence.

**Interpretation.** CoT outputs are longer and appear more rigorous, but they can also introduce extra unsupported assumptions (e.g., implicitly treating "random-looking" as "random") and amplify converse reasoning errors. CoT can therefore increase *explanatory confidence* without improving *logical soundness*. The larger drop in DlgAcc suggests that CoT may especially harm belief-state stability in multi-turn settings.

## 6 Discussion

### 6.1 What CHAOSBENCH-LOGICReveals About LLM Reasoning

**The central finding: weak global constraint satisfaction.** Across all evaluated models, single-question accuracy exceeds 90%, yet multi-turn coherence is far weaker. This indicates that current LLMs can produce correct local judgments but struggle to maintain a consistent global "world model" even when the world model is as simple as a set of boolean predicates linked by Horn-style implications.

**Why scientific domains make this failure more obvious.** In everyday QA, inconsistent answers can be rationalized away as ambiguity. In dynamical systems, however, core relationships are crisp: "random ⇒ non-deterministic" and "chaotic ⇒ deterministic and not random" are definitional in our ontology [19, 14, 3]. CHAOSBENCH-LOGICturns these into explicit constraints that models must respect, revealing when a model's internal reminders are not globally stable.

### 6.2 How CHAOSBENCH-LOGIC Can Support Neuro-Symbolic and Tool-Augmented Systems

The structure of CHAOSBENCH-LOGICmakes it a natural evaluation target for hybrid approaches:

- **Symbolic verifier / constraint checker.** An LLM can propose answers while a symbolic module checks whether the answers violate $\Phi$, triggering self-correction.
- **Logic-guided decoding.** Constrain outputs so that answers cannot jointly violate implications already committed within a dialogue.
- **Training with consistency losses.** Fine-tuning can penalize axiom violations across turns, not only wrong final labels.

Because CHAOSBENCH-LOGICuses a compact and explicit ontology, these methods are implementable without full theorem proving.

### 6.3 Limitations and Scope

CHAOSBENCH-LOGICdeliberately emphasizes *logical structure* over numerical simulation. This makes the benchmark lightweight and auditable, but it also means that we treat each named system as having a canonical regime label. Future releases can extend the benchmark with (i) multiple parameter regimes per system, (ii) items that require extracting regime information from equations or parameter values, and (iii) hybrid items that combine symbolic logic with tool-verified numeric diagnostics (e.g., Lyapunov estimation).

## 7 Conclusion

We introduced CHAOSBENCH-LOGIC, a benchmark for diagnosing logical and symbolic reasoning failures in LLMs within the domain of chaotic dynamical systems. By grounding all questions in an explicit first-order logic ontology, CHAOSBENCH-LOGICmeasures not only per-item accuracy but also implication consistency and multi-turn coherence. Experiments show that frontier and open-source LLMs achieve strong local performance (91–94% per-item accuracy) yet remain globally brittle: dialogue-level accuracy is substantially lower and all tested models fail on compositional items (0%). We hope CHAOSBENCH-LOGICsupports progress toward scientific AI systems that maintain globally consistent beliefs, resist domain misconceptions, and integrate symbolic structure with fluent language understanding.

## References


[1] AI@Meta. 2024. The Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Accessed: 2026-01-04.

[2] Anthropic. 2024. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2026-01-04.

[3] Devaney, R. L. 1989. *An Introduction to Chaotic Dynamical Systems*. Reading, MA: Addison-Wesley, 2 edition.

[4] Gilpin, W. 2021. Chaos as an Interpretable Benchmark for Forecasting and Data-Driven Modelling. *Journal of Physics: Complexity*, 2(4): 045007.

[5] Google DeepMind. 2025. Gemini 2.5 Flash. https://deepmind.google/technologies/gemini/. Accessed: 2026-01-04.

[6] Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021) Datasets and Benchmarks*, 21401–21414. Curran Associates, Inc.

[7] Hénon, M. 1976. A Two-Dimensional Mapping with a Strange Attractor. *Communications in Mathematical Physics*, 50(1): 69–77.

[8] Kuramoto, Y.; and Tsuzuki, I. 1978. Diffuse Instability in Reaction Systems. *Progress of Theoretical Physics*, 59(6): 1752–1773.

[9] Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Association for Computational Linguistics.

[10] Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI 2020)*, 3622–3628. IJCAI Organization.

[11] Lorenz, E. N. 1963. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141.

[12] May, R. M. 1976. Simple Mathematical Models with Very Complicated Dynamics. *Nature*, 261: 459–467.

[13] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

[14] Ott, E. 2002. *Chaos in Dynamical Systems*. Cambridge, UK: Cambridge University Press, 2 edition.

[15] Pathak, J.; Hunt, B.; Girvan, M.; Lu, Z.; and Ott, E. 2018. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Physical Review Letters*, 120(2): 024102.

[16] Rössler, O. E. 1976. An Equation for Continuous Chaos. *Physics Letters A*, 57(5): 397–398.

[17] Sivashinsky, G. I. 1977. Nonlinear Analysis of Hydrodynamic Instability in Laminar Flames—I. Derivation of Basic Equations. *Acta Astronautica*, 4(11): 1177–1206.

[18] Srivastava, A.; Rastogi, A.; Rao, A.; et al. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. arXiv:2206.04615.

[19] Strogatz, S. H. 2015. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Boulder, CO: Westview Press, 2 edition.

[20] Thomas, N. 2026. ChaosBench-Logic: A Benchmark for Logical and Symbolic Reasoning on Chaotic Dynamical Systems. https://github.com/11NOel11/ChaosBench-Logic. Version v1.0. Accessed: 2026-01-04.

[21] Thomas, N. 2026. ChaosBench-Logic Dataset. https://huggingface.co/datasets/11NOel11/ChaosBench-Logic. Version v1.0. Accessed: 2026-01-04.

[22] Uhlenbeck, G. E.; and Ornstein, L. S. 1930. On the Theory of the Brownian Motion. *Physical Review*, 36(5): 823–841.

[23] Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of the Forty-First International Conference on Machine Learning (ICML 2024)*. Vienna, Austria: PMLR.

[24] Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. Curran Associates, Inc.

[25] Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.