# Detecting LLM Hallucination on Dynamical Systems via Physics-Based Audit Signals

**Ahmad A. Rushdi**
Stanford University
rushdi@stanford.edu

## Abstract

When an LLM is asked to numerically solve a dynamical-system problem, its mistakes can be silent: a wrong trajectory looks reasonable, and checking it normally means running the real solver yourself—which defeats the point of asking. We propose two checks that need no reference solution, only the equations of motion the user already knows. The first asks whether quantities that should be conserved (energy, momentum) actually stay constant along the model's output. The second asks whether each step of the output is consistent with the differential equation itself. We test 6 dynamical systems and 6 LLMs and find that each check has a blind spot the other covers, and together they rank-order true error. The checks read only the output trajectory, so they apply whether the model emits numbers directly or runs code it wrote. These checks expose observations about how some LLMs reason numerically: some are reproducible while others vary under identical prompts, and reminding a model about conservation laws can fix some failures while teaching others to game the check.

## 1 Introduction

Scientific pipelines increasingly route parts of their logic through general-purpose Large Language Models (LLMs): drafting simulation code, proposing initial conditions, or producing numerical trajectories directly. This is dangerous because LLM errors on numerical tasks are *silent*: a trajectory that looks reasonable—smooth, bounded, starting from the correct initial condition—can quietly violate energy conservation by orders of magnitude over a single orbit. Unlike factual hallucination ("the capital of France is Cairo"), a physics hallucination carries no lexical signal; a human reading the raw numbers may not notice it without running an independent integrator—a numerically inaccurate trajectory presented as if correct. Text-oriented detectors do not close this gap: semantic-entropy methods [Huang et al., 2023, Farquhar et al., 2024] compare sampled completions for semantic agreement, which does not transfer to trajectories where "meaning" is geometric. Running a reference integrator to audit the LLM defeats the point of calling it; and hallucinations persist because training rewards guessing over admitting uncertainty [Kalai et al., 2025].

We make a simple observation: in physics, priors are free. Hamiltonian systems conserve energy; central-force problems conserve angular momentum; symplectic integrators preserve phase-space volume [Hairer et al., 2006]. These are properties of the governing ODE, available to anyone who knows which equation the LLM was asked to solve. A model that truly integrates the dynamics preserves these invariants; one that pattern-matches will not.

Physics-informed machine learning encodes conservation laws into model architectures or losses [Karniadakis et al., 2021, Raissi et al., 2019, Greydanus et al., 2019, Cranmer et al., 2020], treating invariants as *training* signals baked in before deployment. Agentic frameworks such as Agentic-SciML [Jiang and Karniadakis, 2025] use PDE residuals as *design-time* criteria over candidate architectures. Standard uncertainty quantification [Kendall and Gal, 2017, Kuhn et al., 2023, Liu

Conference on Physics and AI at Stanford University (PAI 2026).

et al., 2025] requires logits, multiple samples, or model internals—none exposed by frontier chat APIs—and assumes outputs have natural-language "paraphrases," which numerical trajectories lack. Scientific LLM benchmarks [Miao et al., 2026, Mohammadzadeh et al., 2025] score correctness *against* ground truth; we ask whether correctness can be *estimated without it*, at inference time, on a single trajectory the model has already produced.

We propose 2 complementary ground-truth-free signals[1]: the *conservation residual* $R$ and *step consistency* $S$—and evaluate them on 6 dynamical systems and 6 LLMs. Together they rank-order trajectory error with Spearman $\rho = 0.85$; we emphasize that the contribution is a *ranking tool*, not a calibrated predictor. Because $R$ and $S$ read only the output, they audit model-written integration code as readily as direct generation; we study the latter as the harder, tool-free regime. Our central empirical contributions are three failure regimes, a deterministic-vs-stochastic split in LLM numerical reasoning, and a double-edged-sword effect of physics-informed prompting.

## 2 Methods: Two Audit Signals

**Task and reference.** Every experiment poses the same request: *numerically integrate a given ODE* $\dot{\mathbf{y}} = f(t, \mathbf{y})$ *from initial condition* $\mathbf{y}_0$ *on* $[t_0, t_f]$ *and return the trajectory at* $N$ *prescribed time points*. Each prompt lists the analytically known conserved quantities $\{\phi_k\}$ (with $\dot{\phi}_k = 0$). The model returns $\hat{\mathbf{y}}_M$ using internal reasoning only—no code, no tools. For *offline validation only*, we compute a reference $\mathbf{y}^\star$ via the DOP853 adaptive Runge–Kutta integrator [Hairer et al., 2006] ($\mathrm{rtol} = 10^{-11}$, $\mathrm{atol} = 10^{-13}$) and define the true normalized error $E = \max_i \|\mathbf{y}_i^\star - \hat{\mathbf{y}}_{M,i}\|_2 / \max_i \|\mathbf{y}_i^\star\|_2$. The central question is whether $E$ can be *estimated without access to* $\mathbf{y}^\star$. We assume the governing equation is known (scientific computing and simulation), so $\{\phi_k\}$ and $f$ are free. Auditing is also cheaper than re-solving: $R$ uses no $f$-evaluations and $S$ exactly $N$, versus an independent adaptive reference—which, where trusted, should simply be used; our signals target settings where re-solution is undesirable or itself untrusted.

**Task** $\mathcal{T} = (f,\ \mathbf{y}_0,\ [t_0, t_f],\ N,\ \{\phi_k\})$

**Given** (JSON prompt): ODE $\dot{\mathbf{y}} = f(t, \mathbf{y})$; initial $t_0, \mathbf{y}_0$; interval $t_f$ and $N$ output points; known conserved quantities $\{\phi_k\}$ ($\dot{\phi}_k = 0$).
**Return** (internal reasoning only — no code, no tools): $\hat{\mathbf{y}}_M$, the $N$ state vectors at the requested times.
**Audit** (post-hoc — no reference solver): $R$ asks whether $\{\phi_k\}$ stay constant along $\hat{\mathbf{y}}_M$; $S$ asks whether each step locally satisfies the ODE.

**Conservation residual $R$.** $R$ measures how much the LLM trajectory violates known invariants, normalized by their initial magnitude:

$$R(\hat{\mathbf{y}}_M) \;=\; \max_k \; \frac{\max_i \; |\phi_k(\hat{\mathbf{y}}_{M,i}) - \phi_k(\hat{\mathbf{y}}_{M,0})|}{\max(|\phi_k(\hat{\mathbf{y}}_{M,0})|,\ \epsilon)}, \tag{1}$$

with floor $\epsilon = 10^{-3}$. Any trajectory violating a Noether-type conservation law must deviate from the true solution; the converse does *not* hold (a trajectory can preserve all invariants while being completely wrong), and this asymmetry is the source of $R$'s blind spots.

**Step consistency $S$.** $S$ checks whether the trajectory locally satisfies the ODE, regardless of invariant preservation:

$$S(\hat{\mathbf{y}}_M) = \max_i \frac{\big\|(\hat{\mathbf{y}}_{M,i+1} - \hat{\mathbf{y}}_{M,i})/\Delta t - f(t_i, \hat{\mathbf{y}}_{M,i})\big\|}{\max\big(\|f(t_i, \hat{\mathbf{y}}_{M,i})\|,\ \epsilon\big)}. \tag{2}$$

$S$ is orthogonal to $R$: a trajectory can have $R = 0$ with large $S$ or vice versa. Both signals require only the LLM output and the system's known ODE.

**Benchmark systems and models.** The six benchmark systems span a difficulty gradient and expose distinct failure modes: the Kepler two-body problem with $e = 0.5$ (capability probe, closed-form solution); the harmonic oscillator (phase-drift probe, since $R$ is invariant under any phase

[1] Code: https://github.com/aarushdi/physics-audit-signals

shift of $\cos t$); the nonlinear pendulum at $\theta_0 = 1$ rad (hybrid); the Chenciner–Montgomery figure-8 three-body orbit [Chenciner and Montgomery, 2000] (chaotic, no closed form); Lotka–Volterra predator–prey [Strogatz, 2015] (non-energy invariant); and a conservative Duffing oscillator [Kovacic and Brennan, 2011] with non-standard parameters ($\ddot{x} = -x - 0.3x^3$, $x_0 = 0.7$) as a novel-ODE probe. We evaluate six frontier LLMs in their most deliberative modes: GPT-5.3, Claude Opus 4.6, Gemini 3.1 Pro, Muse Spark, DeepSeek R1, and Qwen 3. A three-run protocol tests stochasticity (Run 2, identical prompt) and resolution sensitivity (Run 3, halved grid). For evaluation, we use Spearman $\rho$. For every (model, task) pair, we have $R$, $S$, and the hidden true error $E$. We report $\rho$ between a signal and $E$ ($\rho = 1$ perfect ranking, $\approx 0$ unrelated, $< 0$ anti-correlated), rather than Pearson, because the signals span many orders of magnitude and we care about ranking, not absolute calibration.

## 3 Results

**Audit signals rank-order true error.** Figure 1 plots $R$ against $E$ for each task. On Kepler, $R$ rank-orders models well ($\rho_R = 0.83$): models that recognize the analytical solution score $R, E \lesssim 10^{-3}$, while others show substantially larger errors. On the remaining five tasks, $R$ alone provides little ranking power ($\rho_R \leq 0.31$) due to the blind spots described below. Across all 36 pairs, the aggregate Spearman is $\rho_R = 0.69$ (95% bootstrap CI [0.45, 0.84]) excluding 3 $R = 0$ points, dropping to 0.41 when included. Because $S$ checks local ODE satisfaction rather than global invariant preservation, it catches the failure modes $R$ misses: across *all* 36 points—$R = 0$ cases included—$S$ achieves $\rho_S = 0.85$ with no exclusions. $S$ matches or exceeds $R$ on every task and specifically repairs blind spots on harmonic ($\rho_R = 0.09 \rightarrow \rho_S = 1.0$), pendulum ($\rho_R = -0.14 \rightarrow \rho_S = 0.94$), and three-body ($\rho_R = 0.06 \rightarrow \rho_S = 0.53$).

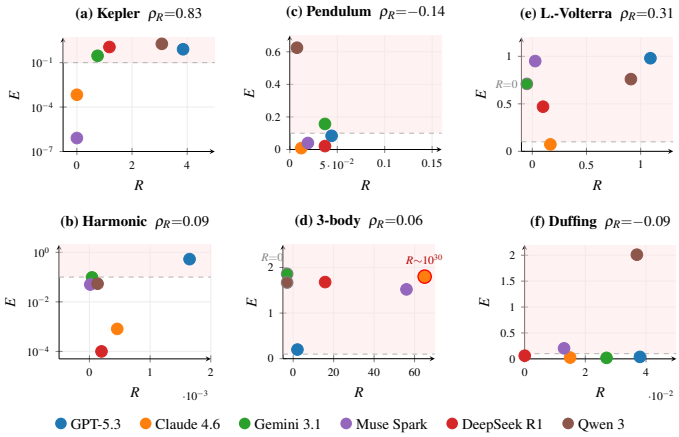
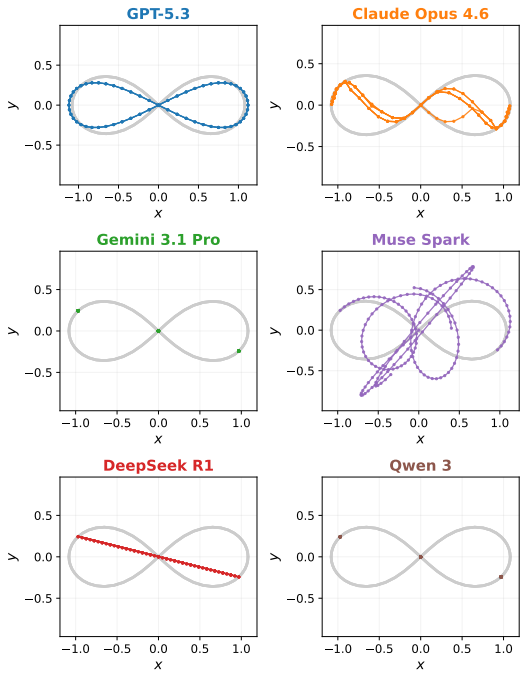


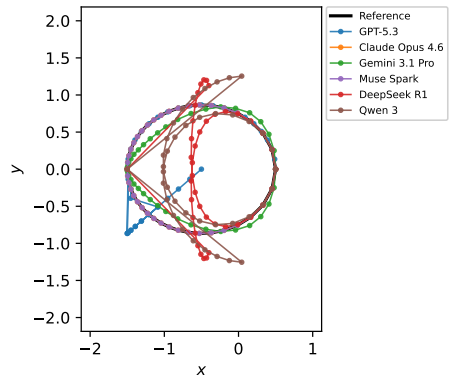

Figure 1: Conservation residual $R$ vs. true trajectory error $E$, per task (one point per model, $n = 6$), with within-task Spearman $\rho_R$ in each title. Pink shading marks $E > 10\%$; scales vary across tasks.

We characterize $R$ and $S$ as *ranking tools*: calibration of $R$ gives in-sample ECE = 1.44 but leave-one-out ECE = 3.14, confirming overfitting at this sample size. Self-reported confidence is no substitute: on the pendulum, Gemini reports 0.85 at $E = 0.16$ while Claude reports 0.25 at $E = 0.008$ [Kalai et al., 2025]. Physics-grounded signals measure the *output*, not the model's opinion.

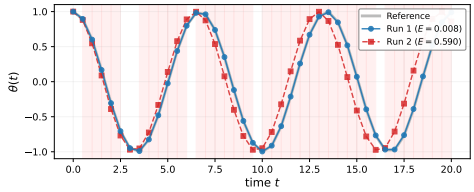
The per-task scatter reveals 3 regimes. **(i) $R$ flags catastrophic failure**: on three-body, Claude produced a trajectory with Hamiltonian drift $R \sim 10^{30}$, the signature of a near-singular close encounter handled incorrectly—a "this trajectory is catastrophically wrong" signal a practitioner can use with no ground truth. **(ii) Trivially-invariant output**: on the same task, Gemini and Qwen emit near-stationary clusters with $R = 0$ but $E = 1.77$ (Fig. 2a)—they satisfy conservation trivially without integrating the dynamics. Noether's theorem guarantees symmetries imply conservation laws, but preserving a conserved quantity does not imply correctness. **(iii) Phase drift on closed-form tasks**: on the harmonic oscillator, all six LLMs claim $x = \cos t$ and preserve energy ($R \sim 10^{-5}$–$10^{-3}$) yet errors span four orders of magnitude (e.g., GPT-5.3 drifts 0.54 rad of phase at $R = 1.6 \times 10^{-3}$). Blind spots (ii) and (iii) are disjoint and precisely addressed by $S$. They are also *visually* diagnostic: in Kepler phase space (Fig. 2b), energy-violating models spiral off the reference ellipse.




(a) Figure-8 orbits: Gemini and Qwen emit near-stationary clusters ($R=0$, $E=1.77$); Claude diverges.




(b) Kepler orbits: energy-violating models spiral or precess off the reference ellipse.




(c) Claude on the pendulum, same prompt: Run 1 ($E=0.008$) vs. Run 2 ($E=0.59$).

Figure 2: Phase-space and stochasticity diagnostics.

## 4 Robustness

Re-running an identical prompt across 20 (model, task) pairs splits models cleanly. Gemini produced byte-identical output on all 5 tasks and Muse on 3 of 5 (*deterministic reasoners*); GPT-5.3 and Claude produced substantially different trajectories every time (*stochastic reasoners*). Figure 2c makes this concrete: same prompt, $E=0.008$ on one Claude run and $0.59$ on the next. Across 7 runs each on Kepler and pendulum, the max/min error ratio ranges from $35\times$ to $1{,}253\times$; on three-body, 5 runs of Claude yield $R$ spanning 31 orders of magnitude—so stochastic reasoners must be sampled several times, not once. Resolution and reasoning-mode ablations are gentler: halving the grid preserves rank-ordering ($\mathrm{Spearman}(R_{41}, R_{21}) = 0.73$), and switching Claude from Extended to standard mode degrades accuracy 13–2687×. Appending "your output MUST preserve [invariant]" to the three-body and Lotka–Volterra prompts has mixed effects. Claude's three-body blow-up vanished ($R: 10^{30} \to 0.41$) and Gemini's trivially-invariant shortcut broke ($R: 0 \to 8.52$), but GPT switched from honest-but-wrong ($R>0$) to trivially-invariant ($R=0$) on both tasks—gaming the invariant without solving the dynamics [Kalai et al., 2025]. The prompt shifts failure modes rather than eliminating them, so $R$ and $S$ should be post-hoc audits, *not* optimization targets shown to the model.

## 5 Conclusion

$R$ and $S$ are not universal detectors: each has characterized blind spots, both require models to fail in distinct ways to rank, and neither calibrates absolute error at this sample size. The headline $\rho_S=0.85$ rests on 36 pairs but is corroborated by 85 extended-mode pairs and the Run 2/3 ablations (§4), with bootstrap CIs reported throughout. Within their domain they are *free*, *black-box*, and more reliable than self-reported confidence: use $R$ where invariants are known and $S$ where the ODE is explicit, treat their disagreement as itself diagnostic, and sample stochastic reasoners several times before trusting any single response. A natural next step is to audit model-written integration *code*, and to extend $R$ and $S$ to PDE discretizations and molecular dynamics—physical invariants as post-hoc *audit signals* against silent numerical hallucination.

## References


Alain Chenciner and Richard Montgomery. A remarkable periodic solution of the three-body problem in the case of equal masses. *Annals of Mathematics*, 152(3):881–901, 2000.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR Deep Differential Equations Workshop*, 2020.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Samuel Greydanus, Mick Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *NeurIPS*, 2019.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, 2nd edition, 2006.

Lei Huang, Weijiang Yu, Weitao Ma, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Qile Jiang and George Em Karniadakis. AgenticSciML: Collaborative multi-agent systems for emergent discovery in scientific machine learning. *arXiv preprint arXiv:2511.07262*, 2025.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *NeurIPS*, 2017.

Ivana Kovacic and Michael J Brennan. *The Duffing Equation: Nonlinear Oscillators and their Behaviour*. Wiley, 2011.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Xiaoou Liu, Tiejin Chen, Longchao Da, et al. Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*, 2025.

Tingjia Miao, Wenkai Jin, Muhua Zhang, Jinxin Tan, Yuelin Hu, Tu Guo, Jiejun Zhang, Yuhan Wang, Wenbo Li, Yinuo Gao, et al. PRL-Bench: A comprehensive benchmark evaluating LLMs' capabilities in frontier physics research. *arXiv preprint arXiv:2604.15411*, 2026.

Saeed Mohammadzadeh, Erfan Hamdi, Joel Shor, and Emma Lejeune. FEM-Bench: A structured scientific reasoning benchmark for evaluating code-generating LLMs. *arXiv preprint arXiv:2512.20732*, 2025.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Steven H Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, 2nd edition, 2015.