WHAT CAN LARGE LANGUAGE MODELS DISCOVER FROM EXPERIMENTS? SCIENTIFIC CLAIMS AND DISCOVERY VALUE ACROSS MULTIPLE WORLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used as scientific agents, yet successful prediction alone cannot establish whether they infer mechanisms or merely fit observations. We investigate this distinction through a controlled multiworld framework for mechanism identification, active experimental discrimination, and conditional deduction beyond observed data. The environment spans ten physics topics and thirty synthetic worlds: worlds within each topic share experimental interfaces but differ in hidden generative rules. Models select experiments, commit to predictions before receiving feedback, and state the assumptions and limits of their conclusions. We analyze 90 final answers from GPT-6, Claude, and Gemini together with their experimental trajectories. Our evaluation separates empirical, mechanistic, and deductive evidence along six claim-level dimensions and relates synthetic rules to established physics, distinguishing mechanistic correspondences, effective approximations, and counterfactual constructions. The results reveal meaningful but bounded achievements. In stochastic motion, calibrated temporal statistics and force interventions constrain hidden dynamics without uniquely identifying them. In lattice systems, closed-path invariants support a complete four-class classification under explicit assumptions, although the apparatus prepares only two classes. Across models, valid local findings coexist with unsupported or falsified stronger explanations. These results support within-task reconstruction and conditional mathematical discovery, but do not establish reliable autonomous discovery or previously unknown physical laws. Evaluating machine scientific discovery therefore requires jointly examining experimental discrimination, generalizable implications, and validity boundaries rather than accuracy or plausibility alone.

1 INTRODUCTION

Large language models are moving from scientific question answering toward agents that formulate hypotheses, write analysis programs, and plan experiments. Interactive environments and automated research systems consequently assess experimental actions, data analysis, and research products (Jansen et al., 2024; Majumder et al., 2025; Lu et al., 2024; Yamada et al., 2025). This raises a more specific question: what scientifically valuable knowledge can a model establish through experiments? Completing a task or improving prediction does not by itself identify the constraints producing a phenomenon, discriminate alternative explanations, or justify generalization. We examine the claims models actually make and the extent to which experiments and arguments support them.

Empirical fitting, mechanism identification, and general deduction can produce superficially similar answers. Multiple mechanisms may explain the same observations, while familiar settings can invite untested textbook explanations. We distinguish observable regularities, mechanistic explanations, and conditional theorems. Their respective support comes from measurements, discriminating interventions, and inspectable proofs. Testing only a textbook-consistent environment provides limited evidence that a model checks when a familiar explanation applies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

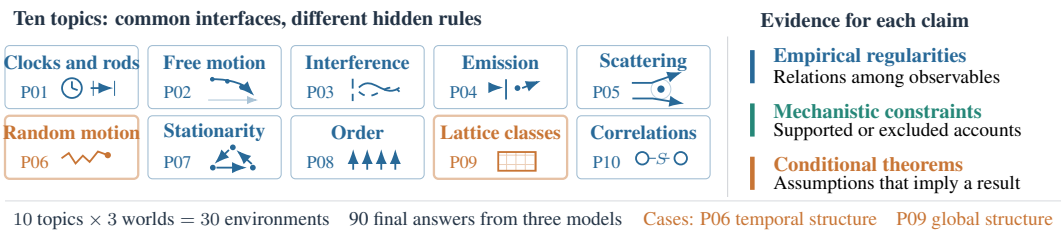


Figure 1: Topic coverage and types of evidence. Orange frames mark P06/P09; each topic contains independent A/B/C worlds. Evidence types are not difficulty levels or model scores. Within-world validity, physical applicability, and human novelty are assessed separately.

Our central design is **the same topic and interface with different hidden mechanisms**. Temporal memory, for example, can arise from active forcing or from an unobserved elastic degree of freedom in equilibrium dynamics. Local lattice records can omit global differences. Models investigate each world independently, while the evaluator compares their claims across worlds. A "world" here is a programmed experimental environment, unrelated to the many-worlds interpretation of quantum mechanics. We also compare each environment with established physics, separating realistic approximations from deliberately counterfactual rules.

Contributions. Our study supports three distinct conclusions about the gap between LLM outputs and scientific discovery.

1. **Models can produce within-task findings beyond fitting.** Useful observables, cross-condition constraints, and conditional classification proofs organize testable knowledge; they do not establish human-unknown laws of nature.
2. **Multiple worlds expose the gap from phenomena to mechanisms.** Similar observations can arise from different rules. A valid local prediction does not validate a stronger explanation; adapting its scope to evidence is more informative than naming a theory.
3. **Scientific value need not require complete recovery.** The three models contribute different observables, fluctuation–response constraints, and conditional deductions. Claim-level evaluation preserves these partial achievements while identifying missing mechanisms and validity tests.

Two retrospectively selected cases illustrate the process, rather than estimate an overall success rate. P06 separates short-time anticorrelation from force-induced motion; P09 extracts invariant information and proves a classification under flat binary periodic-lattice assumptions. Figure 1 situates both cases. One trajectory per model–world and retrospective scoring cannot establish a stable ability ranking. Nor does behavior reveal whether the model recalled training knowledge: we can assess consistency with current evidence and distance from complete physical discovery.

2 Research Questions and Experimental Method

2.1 Discovery targets and the research protocol

Within-world truth is the rule actually generating observations; a **physical reference** is an experimentally and theoretically established relation. Reconstructing a relation absent from the prompt is a within-task finding. Applicability to nature and novelty to humanity require separate verification. We first specify a claim and its conditions, then inspect its observations, counterexamples, or proof.

Each model–world investigation has four rounds and at most two instrument requests per round, with no feedback between the two requests. There are thus at most eight requests, not necessarily eight successful operations. Feedback from rounds 1–3 informs subsequent choices. Round 4 freezes the final claim and predictions before executing its requests; the evaluator checks the resulting feedback, and there is no fifth revision round. Cross-world comparison combines independent trajectories, not a shared model conversation (Figure 2).

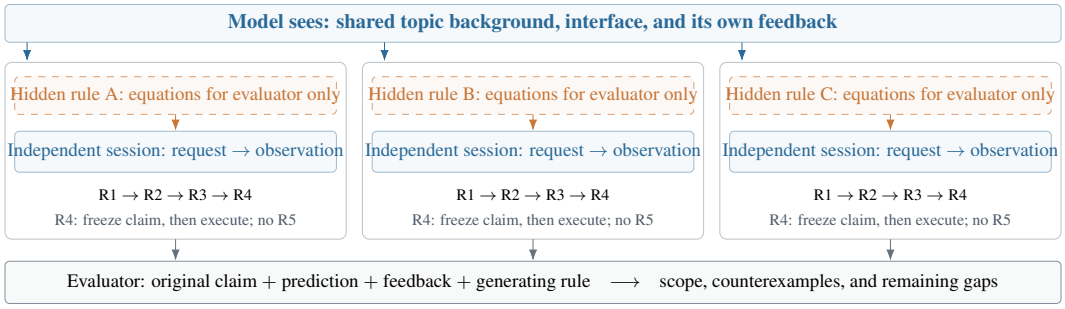


Figure 2: Information flow. Blue is model-visible information; dashed orange boxes denote hidden mechanisms. Sessions do not exchange data. The evaluator reads the generating rules after investigation to check the recorded claims.

Table 1: Six independent questions about a scientific claim. Strong evidence is described rather than converted into an aggregate success rate.

Dimension	Question	Form of stronger evidence
Insight (I)	What actually needs explanation?	Separating confounded quantities that can be measured independently.
Concept (C)	Is a useful object or representation introduced?	Explaining information retained by a statistic but omitted by ordinary summaries.
Explanation (E)	What is explained or excluded?	One structure constraining several observations or a general relation.
Experiment (R)	Can an experiment separate alternatives?	Different predictions stated before a controlled measurement.
Implication (P)	What follows beyond fitted observations?	Supported new-condition predictions or a general proof.
Boundary (B)	Were scope and confounds tested?	Actual changes of conditions and controls, beyond a written caveat.

GPT-6 (request identifier `gpt-6-astra`), Claude, and Gemini each provide thirty final answers; Gemini is archived as AGY. All use the same generator and topic interfaces. Twenty historical prompt branches are excluded. Five GPT-6 trajectories contain continuations from frozen prefixes to fill missing rounds; original failures remain recorded. Appendix A specifies coverage and provenance. Request identifiers do not independently authenticate server-side weights.

2.2 Evaluating conclusions rather than numerical accuracy

Table 1 defines six dimensions. Scores are ordinal: 0 indicates no corresponding achievement or a falsified claim, 1 an unvalidated direction, 2 local support, 3 additional-condition evidence or a relatively complete local argument, and 4 a general result under explicit assumptions. Scores are not summed, and neither answer length nor terminology determines them. Dimension-specific anchors appear in Appendix D.

Valid weak claims and failed strong candidates are assessed separately. Long-scale fluctuation–response agreement cannot rescue a failed short-time diffusion prediction. A falsified explanation and its corresponding implication receive 0, even if its discriminating experiment remains valuable. A conditional proof can receive high explanation and implication scores while untested apparatus assumptions limit its boundary score. Ratings are retrospective, AI-assisted, record-based assessments without blinded domain-expert or inter-rater validation.

3 STOCHASTIC MOTION: DOES MEMORY IMPLY NONEQUILIBRIUM?

3.1 PHYSICAL TARGET AND THREE POSSIBLE WORLDS

The Einstein relation $D = \mu k_B T$ connects long-time diffusion D with weak-force mobility μ; more generally, fluctuation–dissipation relations connect equilibrium fluctuations and linear response (Kubo, 1966). Motion alone is insufficient: one must also measure the response of the same system. Temporal correlation need not imply nonequilibrium because hidden equilibrium variables can retain information about the past.

World A is memoryless thermal diffusion, B adds persistent active forcing, and C couples the observed position elastically to a hidden position in the same-temperature bath. Both A and C represent equilibrium dynamics; B approximates a different physical regime. We follow C because it directly challenges the inference "memory implies active driving or nonequilibrium." A supplies the ordinary-diffusion reference and B the active alternative. Table 2 compares all three. These generating mechanisms are evaluator knowledge, not model input.

The interface allows temperature T, force F, sampling interval, observation duration, and a pinned-probe control. It returns example trajectories and model-specified statistics with standard errors. The pinned probe measures localization error, not the hidden variable. All eight C requests use 6,000 trajectories, interval 0.01, duration 6, and disclosed localization error $\sigma = 0.02$. Units are simulated and $k_B = 1$. Standard errors come from twenty trajectory batches.

3.2 FOUR ROUNDS PROGRESSIVELY CONSTRAIN AN EXPLANATION

Round 1: motion or localization artifact? Writing $M_{\mathrm{obs}}(h)$ for the mean squared displacement over lag h, GPT-6 constructs

$$Q(T) = M_{\mathrm{obs}}(0.2) - 2M_{\mathrm{obs}}(0.1) + 2\sigma^2, \qquad 2\sigma^2 = 0.0008. \tag{1}$$

Independent localization errors add $2\sigma^2$ to each squared displacement. Under stationary, zero-mean increments, the corrected difference equals twice the product expectation of adjacent increments of length 0.1. A negative value therefore indicates anticorrelation. At $T = 1, F = 0$, the free and pinned probes give -0.015922 ± 0.000413 and $0.00000172 \pm 0.00000120$. The near-zero control excludes localization error as a sufficient explanation; it does not identify elasticity or equilibrium.

Round 2: does the correlation scale with temperature? At $T = 0.5$ and 2, Q/T is -0.015579 ± 0.000558 and -0.016462 ± 0.000390, compatible with the original value. Thermal scaling predicts an approximately constant ratio, whereas temperature-independent absolute correlation predicts a changing ratio. This constrains spontaneous motion, but without a corresponding response measurement it is not a complete fluctuation–dissipation test.

Round 3: can anticorrelated motion still respond to force? The model applies $F = \pm 0.2$ and measures $v(F) = \langle X(6) - X(3)\rangle/3$. The normalized responses are 0.5262 ± 0.0415 and 0.4989 ± 0.0418. It removes force-independent drift using the odd response

$$\mu_{\mathrm{win}}(f) = \frac{v(+f) - v(-f)}{2f}, \qquad f > 0. \tag{2}$$

At $f = 0.2$, this gives 0.5125 ± 0.0295. Short-time anticorrelation and positive migration in the measured window coexist, restricting a simple permanently trapped-particle explanation. The subscript emphasizes a finite window, not an established asymptotic mobility or infinitesimal-force limit.

Round 4: a new force magnitude tests the frozen claim. Before feedback, the model states that short-time anticorrelation does not imply absent long-window migration. At unchanged temperature and window it halves the forces, predicting responses near 0.5125. Results are 0.4630 ± 0.0650 and 0.4982 ± 0.0821, giving $\mu_{\mathrm{win}} = 0.4806 \pm 0.0523$, compatible within errors. The reversal and magnitude change constrain drift and some nonlinear alternatives, but two force magnitudes do not establish an infinitesimal linear-response limit. There is no post-feedback model revision. Figure 3 and Appendix F.1 provide the route and all operations.

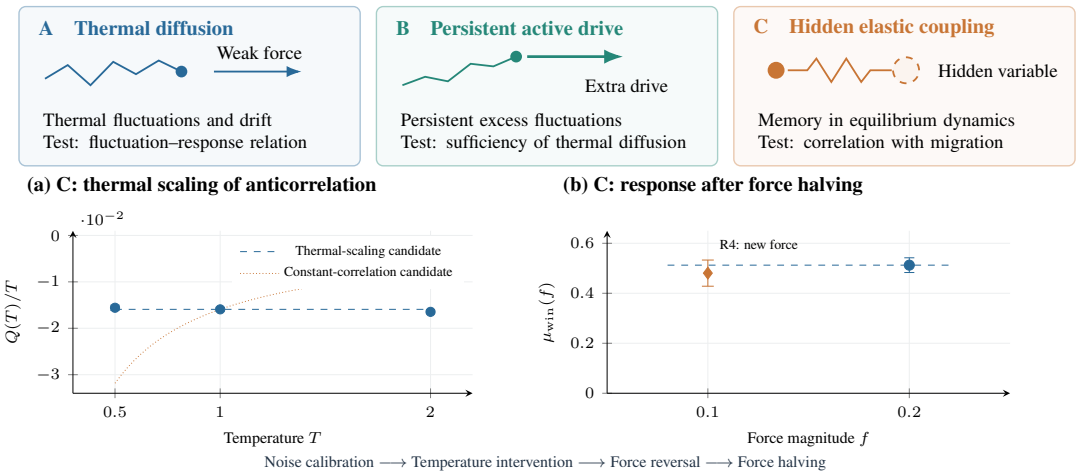


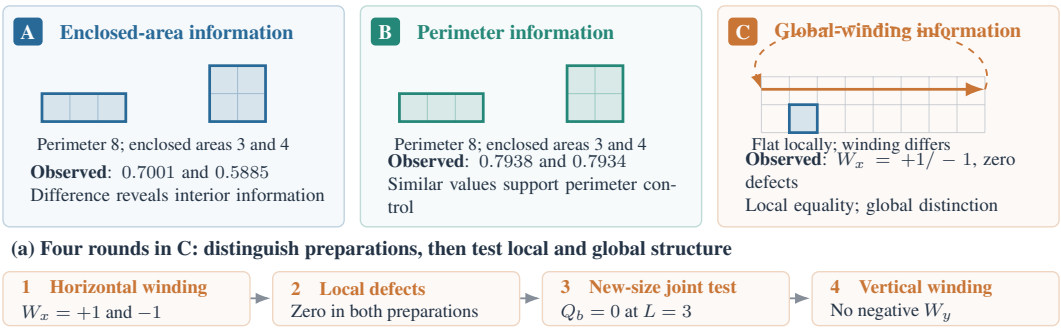
Figure 3: Three stochastic-motion worlds. Top: evaluator-known mechanisms, hidden from the models. Bottom: actual GPT-6 statistics in C and candidate predictions; error bars are one standard error. The dashed response extrapolates the round-3 estimate. Local scaling and migration do not alone establish equilibrium or identify the hidden coupling.

Table 2: P06: generating targets, GPT-6 investigations, and remaining gaps. Truth and physical references are supplied by the evaluator.

W.	Mechanism / physics	Experiments	Observed results	Model conclusion	Correct target	Remaining gap
A	Memoryless thermal diffusion; Brownian approximation.	Pinned control, temperature changes, force at two temperatures.	Corrected variances 0.4978/2.0026; responses 1.007/0.937.	Thermal scaling and compatible response; zero-force statistics do not identify response.	$M(t) = 2Tt$, $\langle \Delta x \rangle = Ft$, $D = \mu T$.	No all-scale memorylessness or complete linear-response test.
B	Persistent active forcing added to thermal motion.	Vary time scale, temperature, and sampling interval.	Correlation differences remain near 0.00563/0.00554.	Relative anomaly can weaken while absolute correlation persists.	Temperature-independent persistent component; thermal diffusion alone is insufficient.	No joint response measurement or unique noise-process recovery.
C	Hidden elastic coupling in an equilibrium bath.	Noise control, temperature change, force reversal and halving.	Nearly constant Q/T; odd responses 0.513/0.481.	Anticorrelation coexists with finite-window migration.	Hidden relaxation; long-time $D = T/2$, $\mu = 1/2$.	No recovered hidden variable, full response kernel, or empirical proof of equilibrium.

3.3 Discovery value and missing evidence

The achievement is to turn an apparent anomaly into separately testable relations. The model also constructs Brownian motion plus an independent stationary restoring process and drift, showing mathematically that anticorrelation can coexist with migration. This proves possibility, not the apparatus's composition. Its final answer explicitly leaves intrinsic memory, nonequilibrium, and asymptotic mobility unconfirmed. Recovering the actual dynamics requires matched-scale fluctuation and response, the hidden coupling or memory kernel, and exclusion of alternatives producing the same finite observations. A, B, and C jointly show why a familiar physical label cannot replace mechanism identification.



(a) Four rounds in C: distinguish preparations, then test local and global structure

Figure 4: Lattice worlds and the C-world evidence chain. Diagrams depict measurement regions; numbers are recorded observations. A/B compare equal perimeters and different areas; C compares local defects with global winding. Each request uses 256 preparations. See Figure 5 for preparation versus conditional classification.

4 Lattices: How Can Locally Identical Records Differ Globally?

4.1 Observable representation and the worlds

Both cases require observables that retain hidden structure: temporal statistics in P06, closed-path products in P09. Here every horizontal or vertical edge of a square lattice carries ± 1. The instrument returns a sample before and after local sign changes, $z_{ij} \mapsto g_i z_{ij} g_j$ with $g_i = \pm 1$. This relabeling changes many raw edges without representing a new physical state. The question is how to extract differences independent of that convention.

The model chooses size L, boundary conditions, a control parameter T, and preparation label $b \in \{0, 1\}$. Each request returns arrays and statistics over 256 preparations. Periodic boundaries join opposite sides, allowing horizontal and vertical winding loops. The label does not reveal which structure changes, and T need not generate thermal excitations in every world. A couples plaquette structure; B generates independent biased edges before relabeling; C is locally flat but permits different global winding. C was selected retrospectively because its trajectory connects redundancy, invariants, local/global distinction, and proof. A/B instead expose area and perimeter information (Figure 4).

4.2 Four rounds identify and test invariant information

Round 1 introduces closed-path products. Every vertex factor occurs twice on a loop, so $g_i^2 = 1$ cancels it. Horizontal winding W_x at $L = 6, T = 0.7$ is $+1$ and -1 for the two preparations. A single edge may reflect notation, whereas this product reads a preparation difference. **Round 2** asks whether that difference already appears locally. At $T = 2$, the model measures the fraction d of plaquettes whose four-edge product is -1; both preparations give zero. "Defect" is this statistical definition, not evidence for a physical defect particle.

Round 3 changes to $L = 3$ and jointly checks local flatness and the expected winding sign using

$$Q_b = d + \frac{1 - (-1)^b W_x}{2}. \tag{3}$$

The nonnegative terms cannot cancel: zero means both requirements hold. Both labels return zero. **Round 4** freezes a conditional classification and tests vertical winding at $L = 3, T = 2$. The statistic $(1 - W_y)/2$ is the frequency of negative vertical winding; both preparations return zero. Each zero summarizes 256 preparations, not an impossibility theorem. The final claim is that identical local closed products can coexist with readable global differences; it does not claim the complete raw arrays are identical.

Table 3: P09: experiments, conclusions, generating targets, and physical limits. Statistics use 256 preparations per request.

W.	Mechanism / physics	Experiments	Observed results	Model conclusion	Correct target	Remaining gap
A	Plaquette interactions; simplified lattice gauge statistics.	Equal-perimeter rectangles of different areas; defects.	Plaquette/winding 0.8859/0.0234; rectangles 0.7001/0.5885.	Perimeter omits interior information.	Loop product equals enclosed plaquette product; independent open-boundary limit yields area law.	No universal area law, defect independence, or full interaction recovery.
B	Independent biased edges with relabeling; statistical control.	Defects, equal-perimeter rectangles, vertical winding.	Rectangles 0.7938/0.7934; winding 0.8594/0.8438.	Local perimeter candidate supported; complete loop invariants classify records.	Simple-loop mean is single-edge mean raised to perimeter.	Finite moments at one size and temperature do not prove independence or uniqueness.
C	Flat global sectors; discrete topological analogy.	Winding, defects, joint test at new size, second direction.	$W_x = \pm 1$; no defects, joint violations, or negative W_y.	Locally equal products can differ globally; conditional classification.	(W_x, W_y) completely classifies flat binary torus records: four classes.	Only two classes prepared; no new quantum phase or full verification of assumptions.

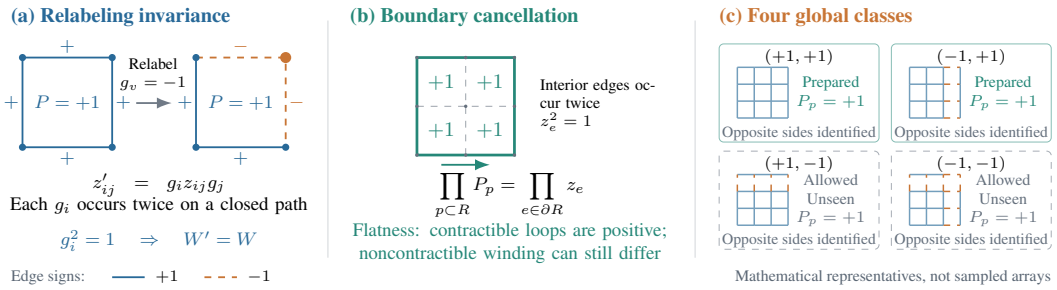


Figure 5: Structure of the P09C classification. Relabeling preserves closed products; multiplying plaquettes cancels interior edges. Equal winding signs suffice for equivalence under the stated flat binary torus assumptions (Appendix F.2). Only the upper two classes were prepared; lower classes are theorem-permitted. Finite-sample zero violations are not the proof.

4.3 From observations to a conditional classification

Are the two winding signs sufficient for every relabeling-invariant difference? "Flat" means every plaquette product is $+1$. Records related by vertex relabeling belong to one class. The model establishes that flat binary records on a connected periodic square lattice without extra punctures have exactly four classes, indexed by (W_x, W_y). Multiplying adjacent plaquettes cancels interior edges; two independent winding directions remain. If two flat records share winding signs, path products of their edgewise ratio construct vertex signs transforming one into the other. Path independence follows because every closed ratio product is 1. This is a sufficiency proof, not extrapolation from zero observed violations (Appendix F.2; Figure 5).

Wilson loops connect local gauge structure to closed observables; A is closest to this idea, while B's independent-edge perimeter law alone says nothing about deconfinement (Wilson, 1974). C relates to global information illustrated by Aharonov–Bohm interference and to $\mathbb{Z}_2$ torus classification, but binary edge records are not continuous electromagnetic phases or quantum states (Aharonov and Bohm, 1959; Tonomura et al., 1986; Kitaev, 2003). The apparatus prepares only the two $W_y = +1$ classes and measures neither a gap, entanglement, nor stability. The conditional theorem is not discovery of a new topological phase.

5 Cross-Model Comparison: Value and Evidential Boundaries

The 90 final answers are evaluated against the same rules and interfaces, although the chosen operations differ. We compare both **the strongest supported claim and the missing step toward the**

Model	Claim being evaluated	I	C	E	R	P	B
GPT-6	P06A: zero-force data do not identify response	3	3	4	3	4	2
Claude	P06A: local diffusion–response relation	3	3	3	3	3	3
Gemini	P06A: diffusion and new-condition displacement	3	2	3	3	3	2
GPT-6	P06B: thermal scaling and persistent correlation	3	3	3	3	3	2
Claude	P06B: thermal and persistent components	3	3	3	3	3	2
Gemini	P06B: superlinear trends and memory candidates	3	2	2	2	1	2
GPT-6	P06C: anticorrelation coexists with migration	3	3	3	2	3	3
Claude	P06C: long-scale relation, short-scale departure	3	3	3	3	2	2
Gemini	P06C: finite-scale sublinear growth	3	2	2	2	1	1
GPT-6	P09C: complete classification of flat records	3	3	4	2	4	2
Claude	P09C: complete classification of flat records	3	3	4	3	4	2
Gemini	P09C: vertex decomposition and global structure	3	3	3	2	3	2

Independent ordinal dimensions: 0 absent/falsified; 1 direction; 2 local; 3 further tests; 4 general result

0 1 2 3 4

Figure 6: Six-dimensional claim scores for the main cases. Colors encode ordinal levels 0–4, not interval-scale quantities; scores are not summed or used for ranking. P06 includes A/B/C; P09 shows C. GPT-6 P06A includes its recorded continuation. High local scores do not certify full recovery; failed stronger candidates appear in Appendix E.

actual mechanism. Figure 6 retains six separate scores. A high score for a narrow theorem does not certify full discovery; Appendix E covers all worlds and failed stronger claims.

P06A/B: connecting fluctuations and response. Claude obtains A-world mobility near 1.045 and a noise statistic of -0.000776, supporting its local Einstein relation. Gemini's short-lag squared displacement and long-window forced displacement also match predictions, without proving all-scale memorylessness. GPT-6's supplemented A responses, 1.007 ± 0.027 and 0.937 ± 0.062, are compatible with one mobility. Its higher implication score instead concerns a counterexample: identical zero-force statistics allow arbitrary force-response functions. In B, GPT-6 separates decreasing relative anomaly from persistent absolute correlation. Claude measures an excess fluctuation of approximately 7.36 beyond the thermal-response contribution; Gemini's heated squared displacement 25.566 falls below both candidate predictions, 28.06 and 30.59. Joint constraints and falsifiable quantitative predictions distinguish these results more than naming active motion.

P06C: valid local relations versus stronger failures. GPT-6 uses noise controls, force reversal, and force halving to support a limited coexistence claim; its boundary score is 3 and discrimination score 2. Claude connects longer-scale fluctuations to response, but the short-lag pure-diffusion candidate predicts 0.022–0.026 against 0.04107 ± 0.00077. Its supported explanation score applies to the longer-scale relation, not this failed extension. Gemini's heated displacement ratio remains about 2.99 versus an earlier 2.95, failing to support its predicted strong weakening of memory. All identify limits of an ordinary all-scale diffusion description; none fully recovers the hidden elastic dynamics.

P09C: a representation versus a complete proof. GPT-6 and Claude provide complete conditional classifications, receiving 4 for explanation and implication. Claude also tests boundary and size changes, with weighted invariant predictions 110 and -91 supported; these are composite statistics, not physical constants. Gemini's vertex/global-factor decomposition captures useful structure but supplies a less complete classification argument. Algebraic factorization does not establish statistical independence of the vertex variables. All lack preparation of all four sectors or quantum-state measurements, limiting boundary scores to 2. Gemini has only two valid rounds here, so proof differences cannot be attributed solely to stable model ability.

Other worlds reinforce this distinction. In P03B, Claude predicts about 20,030 clicks and observes 19,978, whereas GPT-6 mainly recovers visibility and Gemini's readout-collapse prediction fails in this interface. In P04C, all identify a local active band while stronger waiting/reset/energy explanations fail. P07C separates memory from dissipation; P08 motivates symmetry-matched observables and, sometimes, general bounds. One selected trajectory per model–world and retrospective ratings support these claim-level comparisons, not aggregate success rates or a unified winner.

Table 4: Strengths visible in these records and gaps toward complete discovery. These are evidence-based profiles, not population-level ability rankings.

Model	Observed strengths	Important gaps
GPT-6	Useful observables (P06/P08), geometric bounds, observational-equivalence counterexamples, and classification proofs.	Many findings remain local; waiting-time and beam-width predictions fail in P04C/P05C. Stating a boundary does not recover dynamics.
Claude	Joint models of port/count statistics (P03B), fluctuations and response (P06B), and complete classification (P09C).	Unified mechanisms are sometimes premature: absorption, overshoot energy, and soft-core angular formulas face counterexamples.
Gemini	Explicit candidates for affine emission energy, directional/staggered/axial order, and noisy correlation families.	Uniqueness or necessity often exceeds evidence: dispersion-free propagation, readout collapse, and active origins of hidden memory fail or are unsupported.

6 Related Work

ScienceWorld and DiscoveryWorld extend evaluation from language to scientific actions and acquiring unknown knowledge (Wang et al., 2022; Jansen et al., 2024). DiscoveryBench examines structured hypotheses in scientific data, while ScienceAgentBench evaluates scientific programming workflows (Majumder et al., 2025; Chen et al., 2025). We instead compare hidden mechanisms within matched topics and explicitly relate claims to generating rules; this is not a matched-budget benchmark against those systems.

Equation discovery, SINDy, and AI Feynman search compact regularities under different assumptions (Schmidt and Lipson, 2009; Brunton et al., 2016; Udrescu and Tegmark, 2020). Our stochastic-motion case examines the next step: observable selection and interventions distinguishing local statistics from complete dynamics. Discriminating candidate explanations relates to information-based experimental design (Lindley, 1956), although we do not optimize information gain. FunSearch and AI Scientist systems produce checkable constructions or organize automated research (Romera-Paredes et al., 2024; Lu et al., 2024; Yamada et al., 2025); SDE and HEP address discovery evaluation and auditable hypothesis development (Song et al., 2026; Takahara and Mizoguchi, 2026). Our claim is within-task reconstruction, not field-level novelty. Because LLM judges can exhibit verbosity and other biases (Zheng et al., 2023), scores are tied to observations and arguments and remain unvalidated by blinded experts.

7 Conclusion and Discussion

Ten topics, thirty hidden-mechanism worlds, and three models show that LLMs can construct useful observables, design discriminating interventions, and reconstruct constraints under explicit assumptions. The two cases share a central move: ordinary summaries omit relevant structure, so research requires a representation that retains it and experiments or proofs that constrain its interpretation. Temporal anticorrelation and force response provide local dynamical constraints; closed-path invariants support a conditional global classification. These are distinct kinds of achievement.

Multiple worlds separate familiar terminology from actual applicability. Active forcing and equilibrium hidden memory can both generate correlation, while identical local constraints can coexist with different global classes. Across models, valid partial findings coexist with failed stronger explanations. Evaluating claims, supporting evidence, and physical targets separately preserves scientific value without relabeling reconstruction of known rules as discovery of unknown laws. The current records support within-task reconstruction and conditional deduction, not reliable autonomous discovery or human-unknown natural laws. Stronger claims require independent experiments, repeated investigations under controlled conditions, blinded expert assessment, and independent checks of mathematical scope and human novelty. Single trajectories, historical continuations, and retrospective AI-assisted scores delimit the conclusions here.

AI USE STATEMENT

Generative AI assisted development of the multiworld methodology, implementation and inspection of simulator and analysis code, organization of experiments, interpretation of records, development of the claim-level rubric, review of mathematical arguments, literature retrieval, figure creation, manuscript drafting and revision, and translation. Evaluated scientific agents are distinct from AI assistance used in preparing and assessing the study. Observations were generated by fixed simulators, not invented by the tested models. Record checks, continued trajectories, and limitations are documented in the appendices and supplementary archive. AI-assisted review is not independent blinded expert review. Human authors are responsible for the final text, claims, and artifacts; this disclosure does not assert independent expert validation of every result.

REPRODUCIBILITY STATEMENT

The appendices specify the protocol, world-level evidence, scoring anchors, complete case operations, and conditional proofs. The anonymous supplementary archive contains generators, prompts, recorded requests and observations, final answers, continuation provenance, scores, and offline checks. Replaying archived instruments does not require paid model calls. Regenerating language-model answers is a separate operation and need not produce identical text. Original prompts and answers are retained in their original language; manuscript accounts are translated summaries.

ETHICS STATEMENT

Experiments use synthetic physical environments, not human participants. The principal integrity issue is distinguishing program-generated observations and within-task reconstruction from discoveries about nature. We separate model claims, hidden generating rules, translated accounts, and retrospective judgments. AI systems are not authors.

REFERENCES

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your Agent Smarter than a 5th Grader? *EMNLP*, pp. 11279–11298, 2022. https://doi.org/10.18653/v1/2022.emnlp-main.775.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, et al. DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. *NeurIPS, Datasets and Benchmarks Track*, 2024. https://arxiv.org/abs/2406.06769.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, et al. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. *ICLR*, 2025. https://arxiv.org/abs/2407.01725.

Ziru Chen, Shijie Chen, Yuting Ning, et al. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *ICLR*, 2025. https://arxiv.org/abs/2410.05080.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. https://doi.org/10.1126/sciadv.aay2631.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. https://doi.org/10.1038/s41586-023-06924-6.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv:2408.06292*, 2024. https://arxiv.org/abs/2408.06292.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, et al. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv:2504.08066*, 2025. https://arxiv.org/abs/2504.08066.

A. Yu. Kitaev. Fault-tolerant quantum computation by anyons. *Annals of Physics*, 303(1):2–30, 2003. https://doi.org/10.1016/S0003-4916(02)00018-0.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27:986–1005, 1956. https://doi.org/10.1214/aoms/1177728069.

Michael Schmidt and Hod Lipson. Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324:81–85, 2009. https://doi.org/10.1126/science.1165893.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*, 113:3932–3937, 2016. https://doi.org/10.1073/pnas.1517384113.

Roy J. Glauber. The Quantum Theory of Optical Coherence. *Physical Review*, 130:2529–2539, 1963. https://doi.org/10.1103/PhysRev.130.2529.

R. Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29:255–284, 1966. https://doi.org/10.1088/0034-4885/29/1/306.

J. Schnakenberg. Network theory of microscopic and macroscopic behavior of master equation systems. *Reviews of Modern Physics*, 48:571–585, 1976. https://doi.org/10.1103/RevModPhys.48.571.

John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed Experiment to Test Local Hidden-Variable Theories. *Physical Review Letters*, 23:880–884, 1969. https://doi.org/10.1103/PhysRevLett.23.880.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS, Datasets and Benchmarks Track*, 2023. https://arxiv.org/abs/2306.05685.

Zhangde Song, Jieyu Lu, Yuanqi Du, et al. Evaluating Large Language Models in Scientific Discovery. *arXiv:2512.15567v2*, 2026. https://arxiv.org/abs/2512.15567v2.

Izumi Takahara and Teruyasu Mizoguchi. Toward Auditable AI Scientists: A Hypothesis Evolution Protocol for LLM Agents. *arXiv:2607.09195*, 2026. https://arxiv.org/abs/2607.09195.

R. A. Millikan. A Direct Photoelectric Determination of Planck's h. *Physical Review*, 7:355–388, 1916. https://doi.org/10.1103/PhysRev.7.355.

Owen W. Richardson. Thermionic Phenomena and the Laws which Govern Them. *Nobel Lecture*, December 12, 1929. https://www.nobelprize.org/uploads/2018/06/richardson-lecture.pdf.

Richard P. Feynman, Robert B. Leighton, and Matthew Sands. Resonance. In *The Feynman Lectures on Physics*, Vol. I, Chapter 23. Addison-Wesley, 1963. Caltech online edition: https://www.feynmanlectures.caltech.edu/I_23.html.

Kenneth G. Wilson. Confinement of quarks. *Physical Review D*, 10:2445–2459, 1974. https://doi.org/10.1103/PhysRevD.10.2445.

Y. Aharonov and D. Bohm. Significance of Electromagnetic Potentials in the Quantum Theory. *Physical Review*, 115:485–491, 1959. https://doi.org/10.1103/PhysRev.115.485.

Akira Tonomura, Nobuyuki Osakabe, Tsuyoshi Matsuda, et al. Evidence for Aharonov-Bohm effect with magnetic field completely shielded from electron wave. *Physical Review Letters*, 56:792–795, 1986. https://doi.org/10.1103/PhysRevLett.56.792.

Clifford M. Will. The Confrontation between General Relativity and Experiment. *Living Reviews in Relativity*, 17:4, 2014. https://doi.org/10.12942/lrr-2014-4.

E. Rutherford. The scattering of α and β particles by matter and the structure of the atom. *Philosophical Magazine*, 21:669–688, 1911. https://doi.org/10.1080/14786440508637080.

P. G. de Gennes and J. Prost. *The Physics of Liquid Crystals*. Second edition. Oxford University Press, 1993. https://doi.org/10.1093/oso/9780198520245.001.0001.

B. S. Cirel'son. Quantum generalizations of Bell's inequality. *Letters in Mathematical Physics*, 4:93–100, 1980. https://doi.org/10.1007/BF00417500.

A PROTOCOL, RECORD COVERAGE, AND RESEARCH LIMITATIONS

P01–P10 each contain three hidden-mechanism worlds, A/B/C, for thirty primary worlds. Each model independently receives the shared background and interface, chooses operations, reads observations, and proposes explanations and potentially refuting predictions. Neither hidden equations nor other worlds' results are supplied. Four rounds allow at most two requests each, with no intermediate feedback between the two requests. Round 4 freezes the final answer before execution. A final-round counterexample therefore evaluates an already submitted claim; it does not show that the model refused to revise after seeing it. The researcher combines independent trajectories for cross-world comparison.

The GPT-6 request identifier is `gpt-6-astra`; it does not independently authenticate server-side weights. Original collection included thirty primary trajectories and twenty prompt branches sharing the first two rounds, totaling 160 planned calls, 151 valid responses, and nine timeouts. Branches are excluded from the main comparison. Deduplicating shared histories yields 302 instrument requests, 294 successful. Three failures involved integer velocity-array updates; five involved expression syntax or output shape. These counts describe the original collection, not the completed primary corpus.

Continuation of missing rounds. On September 22, 2026, a fixed missingness list specified P02A, P05B, P06A, P09B, and the missing intermediate round of P10A. P02A/P10A continued from round 3 and the others from round 4. All earlier answers and feedback were frozen; no scores, hidden rules, or other-world evidence were supplied. Every suffix from the gap onward was regenerated, so P10A's old final answer remains archived but is excluded from current scoring. Seven new model calls and fourteen instrument requests succeeded. New feedback was replayed with the original seeds and unchanged generator hash. The resulting GPT-6 primary corpus has 120 valid rounds and 240 requests, 232 successful. Historical interface failures remain. Each trajectory still has four rounds and at most eight requests. These separately collected continuations are not independent same-condition replications or evidence of equal computational budgets. Unused prompt branches retain their original missingness.

Table 5: GPT-6 continuations. Answers were saved before their requests executed; original prefixes remain unchanged.

World	Rounds	New test	Feedback and interpretation
P02A	3–4	Infer local propagation scale from clock gradients, then test frequencies.	Positive clock gradient; both delays within prediction bounds; no detected dispersion.
P05B	4	Change charge and beam radius while fixing $q/(ER)$.	Large-angle fractions fail in thin and thick samples, refuting full-distribution compensation.
P06A	4	Apply one force at two temperatures.	Response difference -0.0698 ± 0.0673, compatible with invariant response.
P09B	4	Predict vertical winding from local defects.	0.8594/0.8438, compatible with prediction near 0.846.
P10A	3–4	Repeat endpoints/midpoint and challenge a curved candidate off the old grid.	Amplitudes 0.35373/0.05560 support the linear candidate against the specified alternative.

Claude and Gemini each provide thirty primary trajectories and thirty final answers; AGY denotes Gemini in the archive. Claude has 120 valid rounds and 234 successful instrument requests. Gemini has 112 valid rounds out of 120 and 212 successful requests. Its eight invalid-answer rounds affect seven worlds; P09C has valid answers only in rounds 1 and 4. Neither model was rerun during continuation. The total is 90 final answers, including five continued GPT-6 trajectories. Complete final-answer coverage does not imply all intermediate rounds, instruments, or scientific claims succeeded. Instrument standard errors are not confidence intervals for model ability.

Record-consistency checks passed. Checks cover public tasks, requests, raw and parsed responses, feedback, and scoring evidence. Core generating files match registered hashes; the additional archive's generator and sixty final answers match their source records. This verifies

record/report consistency, not validity of every simulated physical process. Shared interfaces do not imply identical model-selected conditions, so unrelated event counts cannot be compared directly. Invalid parameters, expression errors, and unused controls are not physical counterexamples. Appendix G reports evaluator-run identifiability checks, excluded from model achievements.

The query budget is limited, and some controls do not affect generation. Interpretation applies to implemented time steps, boundaries, and observation channels; energy balances and continuous limits have not been comprehensively verified. Behavioral adaptation cannot reveal whether training knowledge was used. No internal representations, neutral-prompt controls, repeated independent investigations, blinded expert ratings, or inter-rater reliability studies establish stable rankings or human novelty. The anonymous supplement provides generators, public task descriptions, model records, continuation provenance, scores, offline verification, and file hashes. Original responses remain in their original language; English accounts are translations and summaries, with numerical observations unchanged.

B Scientific Targets and Real-World References

Table 6 separates the rules to be reconstructed from established physics. A/B/C are independent worlds, not ordered difficulty levels. More than one world may represent a legitimate physical regime. Targets are evaluator knowledge, not pre-investigation model input. Actual achievements appear in Appendix C; known physical laws are not credited as new discoveries by this study.

Table 6: Ten-topic targets and their correspondence to established physics.

Topic	Correct scientific target	Physical reference and limits
P01	A: common clock/rod scaling and Lorentz-type transformations. B: unchanged clocks/rods and medium-defined propagation. C: common scaling with frequency-dependent signal speed, separate from the kinematic scale.	Vacuum relativity connects dilation, contraction, and Lorentz transformations through invariant c (Will, 2014). A approximates this reference; B resembles low-speed propagation in a medium, not vacuum spacetime; C resembles dispersive probes on a relativistic background. Its speed function is stipulated.
P02	A: acceleration, clock rate, and signal time share a static line element. B: the channels respond independently. C: particles/clocks share geometry but signals have frequency corrections. A position-only potential misses directional velocity effects.	General relativity relates geodesics, proper time, and vacuum light through a metric, with field equations connecting geometry and matter (Will, 2014). A shares the geometric idea but does not solve those field equations. B is a separate-response control; C adds dispersion. None implements complete general relativity.
P03	A/B may share mean fringes, but A has at most one primary click per preparation and B has Poisson counts. C starts from incoherent path mixing, with subsequent device evolution. Complete zero/one/multiple counts distinguish candidates.	Coherent light can interfere with Poisson detection; single-photon preparations can exhibit primary-event exclusivity after loss and dark-count corrections (Glauber, 1963). A/B/C resemble different preparations, not one uniquely real world. Source, detector, and state calibration remain necessary.
P04	A: above threshold, $K = 1.10f - W$; frequency sets event energy and flux controls rate. B: heating changes temperature, which mediates emission. C: resonant accumulation, threshold release, and reset determine waiting times.	Weak-field single-photon $K_{\max} = h\nu - \Phi$ and thermionic $J = A_R T^2 e^{-\Phi/(k_B T)}$ apply in different regimes (Millikan, 1916; Richardson, 1929). A resembles the former; B omits the latter's T^2 prefactor. C borrows classical resonance (Feynman et al., 1963) but stipulates reset; it neither discovers nor refutes photoelectric physics.
P05	A/B deflections depend on impact parameter and q/E, with soft-core corrections in B. C adds many weak vector deflections. Thickness controls encounters; beam radius changes impact parameters. An angular distribution need not uniquely identify microscopic structure.	Rutherford scattering gives $d\sigma/d\Omega \propto E^{-2}\sin^{-4}(\theta/2)$ for nonrelativistic point-Coulomb, approximately single scattering (Rutherford, 1911). A resembles concentrated centers; B finite-scale/screening corrections; C multiple scattering. Effective generators do not calibrate real charge, screening, or nuclear size.

Topic	Correct scientific target	Physical reference and limits
P06	A: thermal fluctuations and weak-force response satisfy an equilibrium relation. B: persistent active forcing adds correlated fluctuations. C: hidden elastic equilibrium variables also create memory. Memory alone does not imply nonequilibrium.	$D = \mu k_B T$ and more general fluctuation–dissipation relations connect equilibrium fluctuations with linear response (Kubo, 1966). A is memoryless, C has hidden equilibrium coupling, and B is actively driven. All have physical analogues; full kernels and real dissipation remain unmeasured.
P07	A: uniform stationary occupancy and detailed balance. B: the same occupancy with nonzero circulating probability flux. C: a reversible microscopic chain projects to history-dependent visible labels.	Detailed balance cancels fluxes pairwise, whereas uniform occupancy alone does not ensure it; coarse observation can induce memory (Schnakenberg, 1976). A/B/C model reversible, driven, and projected processes. Physical entropy production additionally needs local detailed balance, energy exchange, and time calibration.
P08	A: ordinary vector magnitude. B: staggered magnitude with odd-period frustration. C: doubled-angle axial order. A zero ordinary mean need not mean disorder. Finite-size organization does not prove a thermodynamic transition.	Parallel, antiparallel, and head-tail-symmetric order require different order parameters; planar doubled angles represent a restriction of nematic tensor order (de Gennes and Prost, 1993). A/B resemble planar-spin order and C nematic order. Material correspondence, equilibration, and size scaling remain necessary.
P09	A: loops relate to enclosed plaquettes, with area laws under suitable open-boundary conditions and periodic corrections. B: independent edges yield perimeter scaling. C: two winding signs completely classify flat binary torus records into four classes.	Wilson loops and the Aharonov–Bohm phase $q_e\Phi_B/\hbar$ motivate local/global separation (Wilson, 1974; Aharonov and Bohm, 1959; Tonomura et al., 1986). A approximates gauge statistics; B is a statistical control; C is a discrete topological analogy. Four classes are neither universal for gauge systems nor evidence of a new quantum phase.
P10	Under the specified four-setting assumptions, local correlations obey CHSH bound 2, quantum correlations allow $2\sqrt{2}$, and nonsignaling correlations allow 4. Here $S = 4c$. Crossing a bound excludes its hypothesis class, not every implementation.	The CHSH and quantum bounds apply under their stated assumptions (Clauser et al., 1969; Cirel'son, 1980). A is local-compatible, B quantum-realizable, and noiseless C has $S = 3.6$, deliberately beyond quantum theory. Synthetic supraquantum output is not evidence for a natural phenomenon.

C Experiments, Findings, and Gaps in All Thirty Worlds

Each topic begins with the available operations, then compares A/B/C and all three models. Local support, falsified stronger accounts, and untested implications are separated. "Supported" is restricted to the specified conditions; incomplete identification does not mean no scientific value. The main P06/P09 trajectories are detailed in the body and Appendix F. Ratings and failed candidates are consolidated in Appendix E, without counting the same evidence twice.

C.1 P01: Motion, clocks, and propagation

Models vary platform velocity, rod length, signal frequency, and duration, jointly reading clocks, rods, and round-trip times. The question is whether signal speed is also the common scale of clock/rod changes; new velocities, lengths, and instruments test it.

World A. GPT-6 combines clocks, rods, and round trips into a common scaling factor with local extrapolation. This constrains multiple instruments but does not establish a complete relativity principle or unique spacetime implementation; synchronization is specified by the apparatus. Claude derives coordinate transformations and predicts a new negative-velocity clock reading 2.7495, observing 2.74949, with additional coordinate support. Gemini reconstructs the transformation and checks its inverse, obtaining approximately $(1.00090, 0.00087)$ versus $(1, 0)$. These are cross-instrument constraints and conditional algebra, not a complete test of relativity. The source-velocity control does not affect generation and cannot provide physical evidence of source independence.

World B. GPT-6 observes no shared contraction and uses chasing propagation to explain round trips, excluding an A-style scaling account without uniquely identifying an absolute medium. Claude ob-

tains $T = 2Lu/(u^2 - v^2)$ and predicts 2.1976 for new speed/distance settings, observing 2.19790; coordinates distinguish synchronization candidates. Gemini obtains the same form, with two-frequency times 5.33343 and 5.33411. Their adjustment to evidence is valuable, but neither absolute rest in nature nor convention-independent one-way propagation follows from these instruments.

World C. GPT-6 first compares frequencies 1/4 at length 2, obtaining 4.99998/4.04565; at length 0.2 the times are 0.49985/0.40440, restricting fixed startup delay. At speed 0.5, clock/rod ratios 0.86594/0.86585 fit $r(v) = \sqrt{1 - v^2/C^2}$ with $C \simeq 1$, unlike probe speeds 0.8000/0.9887. Its new-speed 0.3 clock prediction 2.86182 matches 2.86188; moving low-frequency round-trip time is 6.15368. It separates scales without recovering full dispersion or transformations. Claude proposes $u(f) = \sqrt{1 - (0.6/f)^2}$; rod/time predictions 1.9079/5.600 match 1.90794/5.59983, while its frequency–wave-number covariance argument retains extra assumptions. Gemini recovers the same dispersion and supports a round-trip prediction near 2.168, but a modified synchronization offset predicts event time 0.5684 versus 0.57733. Correct dispersion does not rescue that failed offset account.

C.2 P02: Free motion and a common structure

The interface provides particle trajectories, accumulated clock readings, and guided-signal times, with controls for initial position/velocity and source shape. A common geometry must connect these channels; one fitted trajectory does not suffice.

World A. GPT-6's first rounds separate velocity-even/odd responses. Continued round 3 measures stationary clocks at $x = 0.43/0.47$, obtaining 2.020831/2.024902 and supporting a positive clock-rate gradient. A proposed $d\tau^2 = A\,dt^2 - B\,d\ell^2$ connects trajectories/clocks to local slowness $n^2 = -\partial_x \ln N/a_{0x}$, $N = \sqrt{A}$. For an unmeasured path of length 0.04 it predicts delay 0.03875. Frequencies 0.7/4 return 0.039126/0.039089, both within 0.00134; their difference 3.74×10^{-5} is below 1.41×10^{-4}. This is local cross-channel closure, not three independent fits, but derivative truncation and spatial variation remain unbounded. Claude's anisotropic quadratic form predicts signal/clock values 0.29065 ± 0.00008 and 0.96620 ± 0.0003, observing 0.290667/0.966181. Gemini's central clock 1.000024 and elliptical-source directional change support a weak-field candidate without fixing the metric. No model derives Einstein field equations.

World B. GPT-6 proposes a single-position clock-gradient/acceleration relation without establishing cross-position or all-instrument geometry. Actual channels respond separately. Claude combines a local quadratic form and an extra coupling; a new-direction time 1.212024 fits 1.21195 ± 0.0002, a useful effective prediction rather than proof of unification. Gemini infers source position from a curved metric and expects a clock substantially above 1.0138, observing 1.003887. Nearby but distinct response functions can create local agreement; it should not receive credit for complete identification.

World C. An integer-array update breaks GPT-6's initial stationary request. After repair, reversed vertical velocities separate effects. Its local candidate is $a = g + A|v|^2 g + B(v \cdot g)v$, with $g \simeq (-0.03883, -0.02016)$, $A \simeq 1.14$, $B \simeq -4.03$. Horizontal reversed velocities give mean acceleration $(-0.02828, -0.02227)$; an oblique difference predicted as $(0.00731, 0.01408)$ is approximately $(0.0076, 0.013)$. Directional extrapolation is useful, but spatial and derivative errors, clocks, and dispersion remain incomplete. Claude changes frequency on a fixed path and predicts $[1.3810, 1.3822]$, observing 1.382022; source-off baseline 1.400020 also fits. Gemini claims absolute frequency independence and predicts 0.787935, but the high-frequency value is 0.786628, about 65 instrument standard deviations away. Geometry-related observations do not justify excluding additional propagation mechanisms.

C.3 P03: Interference and complete event counts

Four-port operations vary path weights, devices, marking, and brightness and report zero/one/multiple-click categories. Mean fringes and event statistics answer different questions: A/B can share first moments with different count distributions.

World A. GPT-6 reconstructs local weight/marking relationships and new-condition mean predictions, without full joint counts or a unique quantum mechanism. The generator applies devices

before path dephasing and allows at most one primary click. Claude introduces a three-component representation but predicts port contrast about -0.59 after full marking and unequal weights through device B. Counts 5917/6044/6092/5996 are nearly uniform, refuting its operation-order explanation. Gemini's hidden-internal-state candidate predicts imbalance under marking changes and a three-device cascade; neither test produces the expected imbalance. Proposed representations and tests remain useful directions, but event exclusivity alone does not identify a microscopic quantum implementation.

World B. GPT-6 contrasts unmarked counts 12050/25/12113/34 with marked counts 6071/5990/6029/6012 and uses single paths to restrict universal-loss accounts. Weights $[1, 4]$ motivate $V = 2\sqrt{w_0 w_1}/(w_0 + w_1)$; $[1, 9]$ predicts 0.6, observed about 0.59, with a partial-marking test. About 5800/30000 multiple-click trials show visibility alone is insufficient. Claude predicts ports $(4230, 1830, 4230, 1830)$ and about 20030 zero-click trials, observing $(4266, 1873, 4275, 1859)$ and 19978. Its read-record control retains contrast near 0.5, supporting weight/marking and Poisson-count relations. Gemini predicts reading the record will collapse two bright ports into four uniform ones; actual counts $(9701, 32, 9550, 30)$ disagree. The read switch is unused in generation, so this is not new physics of quantum measurement.

World C. After near-uniform mean counts, GPT-6 varies brightness 0.1/0.5 and obtains 27448/17920 zero-click trials out of 30000. The candidate $(1 - 0.8b)e^{-0.004}$ fits rates 0.91493/0.59733; specified common-dead-time Poisson predictions 0.8480/0.4454 do not. At brightness 0.25, read/unread zero-click counts are 23878/23951 and multiple clicks 19/16. This constrains primary-event exclusivity but cannot localize it uniquely to source, propagation, or detector. Claude claims cross terms incompatible with nonnegative single-path mixing and predicts shares $(0.081, 0.419, 0.081, 0.419)$, observing approximately $(0.036, 0.453, 0.041, 0.469)$. Repeated path labels $[1, 1]$ only open path 1, so this is not a two-distinct-path control. Gemini's B/C device tests remain near uniform and do not establish anticipated interference control. Neither resolves the distinction between initial mixing and coherent superposition.

C.4 P04: Emission, energy, temperature, and internal state

Models control frequency, intensity, duration, temperature, and material, reading counts, early event times, and energy samples. Frequency selection, waiting times, and event energies are evaluated separately.

World A. GPT-6 separates event energy from count and fits $K \simeq 1.10f - 1.15$: no emission at $f = 1$ and $K \simeq 0.17$ at $f = 1.2$ are supported. This does not uniquely establish discrete delivery or all material responses. Claude changes to material B and observes energies 1.8/0.7, difference 1.1, supporting a common slope; count 126 exceeds its 72–114 prediction, so the complete yield model is not validated. Gemini predicts energy 2.15 at $f = 3$, while low-frequency heating still produces no event. Its uniqueness language exceeds finite-frequency evidence. The synthetic affine slope is not a newly measured Planck constant.

World B. GPT-6's clamped/free-heating counts 2/320 and fixed-high-temperature flux counts 136/150 identify temperature as a mediator without reconstructing all absorption, cooling, or activation terms. Claude also identifies heating but predicts final temperature 3.07 ± 0.2 at high frequency, observing 7.06997; frequency-independent absorption fails. Gemini's low-frequency repeated heating gives 2.76911, below approximately 6.2 at its earlier high-frequency condition, supporting frequency-dependent absorption. Temperature-mediated emission does not imply that frequency cannot affect heating. Neither recovers every factor of real thermionic emission.

World C. GPT-6 obtains matched middle/high-frequency counts 43/0. Equal-total-input runs yield 26/17 events and intervals 0.452143/0.279643, excluding fixed-dose or fixed-time scaling. At new intensity 1, the interval 0.640714 supports a weak directional claim but falls outside the frozen additive prediction 0.7396 and band $[0.67, 0.81]$. A matched repeat returns the same 17 events and prefix, checking deterministic reproducibility rather than adding independent random evidence. Claude's accumulation/threshold/overshoot account predicts count ranges containing 23/47, but energies 0.000649/0.000546 are outside its bands; first event 0.67 also slightly exceeds 0.65. Gemini's heated control retains spacing 0.5, but a pulsed test produces only one initial event instead of ten under complete reset. The true damped oscillator retains phase/internal-state dynamics during dark

periods. These tests restrict specific timing, energy, and reset hypotheses, not every accumulation model. Small deterministic energy samples do not establish quantized escape energy.

C.5 P05: Scattering statistics and internal structure

Each request launches 30000 probes. Energy, charge, thickness, and beam radius vary; angular bins are observed. Encounter probability and conditional deflection must be separated because thickness controls opportunities and beam geometry changes impact parameters.

World A. GPT-6 obtains first-bin counts 27236/11010 at thicknesses 0.1/1, then varies charge/energy and beam radius 2/5. Near-straight fraction is about 0.904; tails above 0.7 are about 0.04747/0.00683. New combinations fixing $q/(ER) = 0.5$ have similar principal bins. This constrains encounters versus deflection without uniquely recovering potential or nuclear size. Claude proposes thickness and $q/(2ER)$ scaling; a new-condition tail count 6880 lies in 6670–7220, with support in another combination. Gemini supports q/E scaling and an energy-loss extrapolation 0.06 but interprets a beam-induced impact-parameter cutoff as a physical screening radius. Effective scaling is better supported than unique microscopic structure.

World B. GPT-6 compares $(E, q) = (2, 2)$ and $(1, 1)$, obtaining 2146/2150 counts above 0.35. Wider beams preserve the near-straight fraction but alter conditional angles. Continued round 4 changes $R = 3$ to 5 and $q = 1$ to $5/3$, preserving $q/(ER)$. Thin/thick large-angle predictions 0.00150/0.04480 fail against $470/30000 = 0.01567$ and $3467/30000 = 0.11557$. The valid discovery is separation of opportunities and conditional scattering plus localization of a geometric effect; the failed scaling does not cancel local q/E support. Claude's soft-core formula predicts fewer than 30 events in three small-angle bins, observing $5 + 22 + 62 = 89$, and misses other bin proportions. Gemini notes separately controlled encounter, loss, and angle statistics; loss 0.06 and vanishing large-angle counts at tiny q/E fit directions, but do not uniquely establish an exact microscopic implementation or absence of physical cross sections.

World C. GPT-6 varies thickness and energy and fixes $q\sqrt{t}/E$ to test distribution shape, finding one scale insufficient. At $(E, q, t) = (1, 1, 0.25)$, radii 2/5 give first-bin counts 5820/17606 and tails above 0.35 of 1416/149, refuting radius invariance. Claude's compound-Poisson candidate locally fits first-bin count 17249 near a prediction 16800 and some tail trends; another second-bin count 6547 exceeds $4000 \pm 30\%$. Gemini expects single-digit counts in $[0.7, 1.2)$ under a soft-core cutoff, observing 47, and misses the many-weak-center accumulation mechanism. Tail sensitivity has value, but complete distributions and unique microscopic recovery remain unestablished.

C.6 P06: Stochastic motion, memory, and response

The interface provides 1000–6000 trajectories, custom statistics, and standard errors, with temperature, force, sampling, and pinned-probe controls. The main C trajectory appears in Section 3; all requests and hidden dynamics appear in Appendix F.1.

World A. After repairing output shape, GPT-6 obtains adjacent-increment products 0.00365 ± 0.00532 (free) and -0.000400 (pinned); the difference is not significantly nonzero. Temperature 0.5/2 yields corrected variances 0.49784/2.00261. Continued round 4 applies $F = 0.2$ for duration 6 and measures $r(T) = \langle X(6) - X(0) \rangle / (6F)$, obtaining 1.0067 ± 0.0271 and 0.9369 ± 0.0616. Difference -0.0698 ± 0.0673 does not cross the prerequest three-standard-error threshold. Zero-force drift, linearity, and all scales remain untested. Its general counterexample sets $Y(t) = \mu(T)Ft + Z_T(t)$ with independent stationary increments of variance $2aT\Delta t$: arbitrary $\mu(T)$ gives identical zero-force statistics. Thus thermal variance scaling cannot uniquely imply the Einstein relation. The E/P scores of 4 concern this counterexample. Claude predicts a localization statistic near -0.0008, observing -0.000776, and obtains mobility 1.045 under reversed force. Gemini's squared displacement 0.4136 ± 0.0075 fits approximately 0.4048, and forced displacement 1.1467 ± 0.0523 fits 1.212. None uniquely establishes an all-scale memoryless Newtonian medium.

World B. GPT-6's normalized anomalies at low/high temperature are 1.4817/1.1237, while differences between separated-lag covariances remain 0.00563/0.00554; round 4 changes sampling. A weaker relative anomaly can coexist with persistent absolute correlation. Claude combines heated zero-force variance and forced displacement, obtaining $31.305 - 20 \times 1.1973 \simeq 7.36$ within its 3.2–10 prediction, supporting thermal plus persistent components. Gemini's heated squared dis-

placement 25.566 falls below nonthermal/thermal candidate predictions 28.06/30.59. The former makes a more direct joint constraint; neither uniquely reconstructs the noise process or a universal superdiffusive exponent.

World C. GPT-6 measures corrected $Q = -0.01592 \pm 0.00041$ versus near-zero pinned control and normalized temperature values $-0.01558/-0.01646$. Reversed forces ± 0.2 yield $v/F = 0.5262 \pm 0.0415$ and 0.4989 ± 0.0418; final half-forces give 0.4630 ± 0.0650 and 0.4982 ± 0.0821. Anticorrelation and migration coexist, without full kernel identification. Claude finds longer-scale equilibrium-compatible relations but rejects its short-lag diffusion candidate: prediction 0.022–0.026 versus 0.04107 ± 0.00077. Gemini's heated ratio $4.9553/1.6589 = 2.987$ remains near 2.95, failing to support its predicted weakening of memory. Scale dependence is a meaningful finding, not proof of nonequilibrium or a thermodynamic anomalous exponent.

C.7 P07: Uniform stationarity, currents, and hidden memory

Models inspect one-step transitions, triples, and dwell histories of three visible labels, vary drive, or request custom statistics. Uniform occupancy, detailed balance, and sufficiency of the current label are different properties.

World A. GPT-6 varies drive/duration, obtains corrected final statistics 0.32745/0.33256, and constructs an effectively reversible visible process. Adding an independent hidden driven process leaves all label statistics unchanged, proving visible reversibility need not identify full-system reversibility. Claude's final circulation statistics $-0.00413/0.00517$ meet its bounds; it also constructs an invisible irreversible process. Gemini predicts a two-step label difference near 26400, observes 26338, and obtains label sum 30174 consistent with uniform occupancy. Energy and bath information are still needed for thermodynamic entropy production.

World B. GPT-6 identifies uniform occupancy with circulating flux and predicts a three-step directional endpoint statistic -0.01125, observing -0.01207 and -0.01037 on repeat. A weak reversed endpoint bias can arise by composing the same one-step chain; it is not reversed forcing. Under circular symmetry and first-order Markov assumptions, Claude derives $R_2 = s^2 + (1-s)^2/2 - J^2/2$, with $s = P_{ii}$, $J = P_{i,i+1} - P_{i,i-1}$, and $R_2 = \Pr(S_{t+2} = S_t)$. At new drive it measures $J \simeq 0.2204$ and three-step return 0.3262 within 0.318–0.333. Gemini reconstructs $P_{ii} = 0.4$ and $P_{i,i\pm1} = 0.3 \pm 0.15D$, matching the drive-dependent generator. Its expectation that changing lag returns P^2 fails because the control is unused; that interface issue does not refute the matching one-step relation.

World C. GPT-6 repairs an expression and observes history statistics 0.04603/0.03141 in round 3 and 0.07943/0.04921 for a different statistic in round 4. These cannot be compared as the same measure. History adds prediction beyond the current label. Projection of a reversible six-state chain supplies a counterexample to "memory implies irreversibility," without uniquely determining hidden state count/matrix. Claude compares continued residence immediately after entry versus generic times, about 0.69/0.79, excluding visible first-order Markov dynamics. Reversal statistics 0.00103/0.00147 fit predictions, while it notes that zero measured asymmetry does not prove process-wide reversibility. Gemini attributes memory to active driving, but the actual symmetric six-state chain is drive-insensitive; one request at drive 5 is out of range. History dependence is an observation, not support for that active cause.

C.8 P08: Observables that identify organization

Models observe planar arrows on a three-dimensional periodic lattice, varying temperature, size, field, and sampling steps. Each result aggregates 24 chains. Parallel, staggered, and axial order require appropriate statistics, not an ordinary mean alone.

World A. GPT-6 obtains squared vector magnitudes 0.8983/0.0911 at low/high temperature for size 4, and 0.8921/0.0522 for size 5. Fixed projections and weak fields distinguish global rotation from internal organization. A vector identity bounds the fraction of pairs separated by at least 90°; the low-temperature size-5 estimate gives about 10.88%, with sampling uncertainty. Claude predicts weak-field mean cosine above 0.8 and obtains final 0.89960 ± 0.00915, but its half-run value 0.77049 differs by 0.12911, failing its simultaneous stability threshold 0.05. Gemini's new-

temperature magnitude 0.5000 and field value 0.3752 fit its ranges; stronger O(2) universality and critical-temperature claims lack finite-size scaling and equilibration evidence.

World B. GPT-6's ordinary squared means are only 0.00047/0.00826, but staggered values are 0.8959/0.0978. Odd sizes 3/5 give neighbor correlations $-0.3772/-0.7092$, motivating the conditional bound $C_1 \geq -\cos(\pi/L)$. Final rotation and twist-deviation checks do not uniquely identify a uniform helix or infinite-size transition. Claude's same-chirality uniform-twist candidate predicts above 0.5 but returns 0.1942 ± 0.0675; its previously listed independently chosen axis-chirality branch, 0.2–0.45 with large error, remains compatible. High-temperature staggered square 0.01532 misses 0.04–0.08. Gemini's even-size neighbor cosine -0.9174 supports a prediction below -0.85. Strong finite-size antialignment does not prove thermodynamic long-range order or critical behavior.

World C. GPT-6 first obtains small ordinary/staggered differences 0.0070/0.0010, then defines $M_2^2 = |n^{-1}\sum_j e^{2i\theta_j}|^2$, obtaining 0.9031/0.0912 at low/high temperature. Fewer steps and larger size retain low-temperature values 0.8986/0.8943. Final first-harmonic terms sum to about 0.0341; new-size high-temperature doubled-angle value is 0.0379. Common-axis structure is supported without a unique energy function or critical point. Claude predicts high-temperature doubled-angle magnitude 0.08–0.35, observing 0.2231. First-harmonic field response 0.0463 is slightly below 0.05 but has error 0.0121, insufficient for strong rejection; ergodicity remains open. Gemini obtains doubled-angle magnitude 0.7844, but ordinary magnitude under uniform field is 0.0735, not its strong-polarization branch above 0.5. Claude also supplies a complete cancellation proof for a conditional equilibrium-symmetry response theorem, supporting E/P scores of 4; unverified equilibrium, local symmetry, and ergodicity limit boundary score to 1.

C.9 P09: Closed invariants and local/global distinctions

The instrument returns edge arrays before/after relabeling and aggregates closed statistics over 256 preparations. Size, boundary, and control parameters vary. C's principal trajectory is in the body; its full proof is in Appendix F.2.

World A. GPT-6 observes plaquette/winding means 0.8859/0.0234. Equal-perimeter-8 rectangles of areas 3/4 give 0.7001/0.5885, revealing interior information. The loop/plaquette product identity is general, but periodic constraints preclude blindly invoking independent-plaquette area laws. Claude retains the periodic total-product constraint and derives $\langle W\rangle = (t^A + t^{N-A})/(1+t^N)$, where $t = \tanh(1/T)$ and $N = L^2$. Prediction about 0.952 matches 0.9566 ± 0.0083. Gemini extrapolates local independence to periodic systems and invokes a quantum toric-code thermal state; its final open-boundary corner/interior means 0.9063/0.8984 do not show the anticipated sharp difference. Handling the global constraint, not naming gauge theory, is the relevant distinction.

World B. GPT-6 estimates $q = 1 - 2\rho = 0.89453$ from defects and proposes $x_{uv} = s_u s_v \eta_{uv}$ with independent biased edges. It predicts $q^2 = 0.80019$ for two perimeter-8 rectangles, observing 0.79340/0.79384. Continued vertical-winding prediction $q^{3/2} \simeq 0.846$ matches 0.85938 ± 0.03202 and 0.84375 ± 0.03361. It also proves that all plaquette products plus two fundamental winding products completely specify relabeling classes for general binary periodic lattices, using edge ratios and a spanning tree. Finite moments at one size/temperature do not prove edge independence or a quantum phase. Claude's new rectangle 0.4891 ± 0.0169 fits 0.44–0.52; winding 0.6341 fits 0.58–0.74, without full temperature dependence. Gemini studies boundary/size effects but does not use equal-perimeter/different-area controls to identify independent-edge generation. Gauge invariance alone cannot determine a Hamiltonian.

World C. GPT-6 measures horizontal winding, defects, a new-size joint statistic, and vertical winding, obtaining $W_x = \pm 1$ and zero violations without negative W_y. Path-independent vertex signs prove four conditional classes. Claude explicitly states completeness of (H_x, H_y), reducing loops to winding and constructing relabeling signs for the ratio of two records. Final composite predictions $110/-91$ both match. Neither establishes that the instrument prepares all four classes or uniquely identifies its algorithm. Gemini proposes $e_{ij} = \sigma_i \sigma_j \eta_{ij}^{\mathrm{top}}$ and a discrete lemma, but the representation does not entail independent σ_i. A zero-compatible two-edge statistic cannot prove joint independence and changes under relabeling. With only two valid rounds, its less developed proof is not grounds for stable ability ranking.

C.10 P10: TURNING EXPLANATIONS INTO TESTABLE BOUNDS

Two settings at each endpoint produce ± 1 outputs. Full counts and random replacement noise are available. With product means E_{xy}, define $S = E_{00} + E_{01} + E_{10} - E_{11}$. Models must translate local-common-cause assumptions into constraints.

World A. Continued GPT-6 repeats zero noise and a midpoint; maximum probability residuals 0.01747/0.01240 remain below frozen thresholds. It proves that $P_v(a, b \mid x, y) = [1+vab(-1)^{xy}]/4$ admits setting-independent local realization iff $|v| \leq 1/2$: CHSH gives necessity; mixtures of four local product tables, a shared fair sign, and independent noise give sufficiency. It constructs a curved noise-response candidate matching every previous grid point, then tests noise 0.125/0.875. Linear predictions 0.34993/0.04993 match 0.35373/0.05560 within 0.0468, unlike curved predictions $0.43/-0.03$. Finite-grid agreement alone cannot establish continuum linearity. Claude also proves locality iff $|v| \leq 1/2$ by eight deterministic strategies; correlation 0.3989 and noisy CHSH 1.2168 support predictions near 0.4/1.2. Gemini finds linear attenuation but makes no new final-round requests. Existence of a local realization never uniquely identifies the actual causal implementation.

World B. GPT-6 tests noise 0/0.5, complete replacement/repeated endpoint, 0.1/0.35, then 0.35/0.23, using 30000 trials per request and 7500 per setting. It derives the local bound and predicts a crossing near noise 0.2312. This is crossing a distributional class boundary, not a physical phase transition. Claude predicts CHSH about 2.028 at noise 0.22 and observes 2.03173. Gemini predicts approximately 2.329 at noise 0.1 and observes 2.34267. They exclude the specified local class while leaving many quantum-realizable implementations. Calling direct conditional joint sampling "superdeterminism" is unwarranted: simulator implementation does not uniquely determine an experimental causal account.

World C. GPT-6 uses endpoints and noise 0.4/0.48 to fix $v \simeq 0.89613$, writes $c = v(1-q)$ and $P(a, b \mid x, y, q) = [1 + ab(-1)^{xy}c]/4$, and tests 0.25/0.75. For nonnegative c, locality holds iff $c \leq 1/2$; $4c \leq 2(1-w) + 4w$ and a saturating mixture imply minimum nonlocal fraction $\max(0, 2c-1)$. Claude instead defines $f = 1 - \sum_\lambda \min_{xy} q(\lambda \mid xy)$, the hidden-distribution mass not shared across settings, and proves its minimum $\max(0, 2v_{\text{eff}} - 1)$ with $v_{\text{eff}} = S/4$. At noise 0.4, $S = 2.16027$ fits 2.09–2.22. This specified dependence measure is not automatically a physical fraction of nonlocal events or universal across dependence measures. Gemini identifies a noisy PR-type table near visibility 0.9 and predicts correlation 0.81, observing 0.81453, without proving a unique distribution or sampler. All are within-task reconstructions of known constraints, not natural supraquantum discoveries.

D CLAIM-LEVEL SCORING RULES AND EXAMPLES

Evaluation first identifies a full sentence as the target. "The tested middle frequency is more active" is local; "every intensity's waiting interval is recovery time plus replenishment time" asserts a broader mechanism. Conditions, supporting records, and counterexamples are recorded before filling six columns. Eloquence is not a substitute for evidence, and score dispersion is not a reason to change an anchor.

Insight and concept concern how the problem is represented. Naming a theory or suggesting an unmeasured direction supports at most 1. A computable statistic with demonstrated local use supports 2; connecting several observations and explaining why conventional readings omit information supports 3. A general necessity/sufficiency argument or systematic construction is needed for 4. Doubled-angle statistics have a defined head-tail equivalence interpretation; unexplained complicated expressions do not receive equivalent credit.

Explanation and implication concern validity scope. Fitting parameter-setting observations is local compatibility. New-condition tests provide additional support; a theorem requires assumptions and an inspectable proof. P04's two-point waiting formula receives 0 for the corresponding explanation/implication when it fails at a third intensity. P09's valid conditional classification can receive 4, without credit for unimplemented quantum phases. A weaker correct claim never cancels a counterexample to a stronger one.

Discrimination concerns the reason for a measurement and whether distinct candidate predictions precede it. Unexecuted suggestions provide limited evidence; completed controlled tests separating

candidates can score higher. Refuting a model's candidate does not erase the experiment's value. Boundary scores require actual condition changes or controls, not a generic promise of more work. Missing preparations, few sizes, or absent independent noise calibration limit the score.

These are ordinal evidence levels, not a calibrated interval scale or an aggregate discovery rate. Equal evidence may receive equal scores. Ratings use saved final answers and actual feedback, with frozen-prefix continuations documented separately. General bounds, saturating constructions, classifications, and nonidentifiability counterexamples face the same standard across models; algebraically rewriting a fit alone is insufficient. Assessments are retrospective and AI-assisted without blinded domain experts or inter-rater validation. No stable model ranking is asserted.

E Unified Claim Scores across All Worlds

Table 7 covers all ninety final answers. GPT-6 P02A, P05B, P06A, P09B, and P10A include actual continuations; other ratings are unchanged. Vectors follow insight/concept/explanation/discrimination/implication/boundary (I/C/E/R/P/B), each 0–4, without summation. Claims of different scope cannot be ranked by their vectors alone. Evidence and remaining gaps appear in Appendix C.

Table 7: Three-model claim scores in all thirty worlds. Stronger failed candidates are listed separately.

World	GPT-6	Claude	Gemini
P01A	**2/2/3/2/3/2** Shared clock/rod factor and round-trip cancellation.	**3/3/3/3/3/2** Common structure of clocks, rods, and synchronization.	**2/2/3/2/3/2** Lorentz-type transformation and inverse mapping.
P01B	**3/2/3/3/3/2** Unchanged clocks/rods and chasing propagation.	**3/3/3/3/3/2** Unchanged clocks/rods and chasing propagation.	**3/2/3/3/3/2** Absolute time and round-trip speed relation.
P01C	**3/3/3/3/3/2** Separating two speed scales.	**3/3/3/3/3/2** Dispersive signals versus the clock/rod scale.	**3/3/3/3/3/2** Dispersion separated from clock/rod scaling.
P02A	**3/3/3/3/3/1** Velocity-even response and local trajectory/clock/signal closure.	**3/3/3/3/3/1** Local quadratic form connecting three channels.	**2/2/2/2/2/1** Weak-field geometry and directional response.
P02B	**2/2/2/2/2/1** Single-point clock-gradient/acceleration correspondence.	**3/3/2/3/3/1** Effective prediction from a local quadratic form.	**2/2/2/2/2/1** Source-dependent clock and path response.
P02C	**3/2/3/3/2/1** Directional cross term in initial acceleration.	**3/3/3/3/3/2** Geometry plus signal-frequency corrections.	**2/2/2/3/2/1** Geometric candidate for velocity-dependent response.
P03A	**2/2/2/2/3/1** Weight/marking dependence of port contrast.	**3/3/0/3/0/1** Candidate: dephase first, then reconstruct contrast with a device.	**1/1/1/2/0/1** Hidden internal state and broken-cascade candidate.
P03B	**2/2/2/2/3/1** Local weight-dependent visibility.	**3/3/3/3/3/2** Mean relations plus Poisson counts.	**2/2/3/3/3/1** Different port evolution under B versus BA.
P03C	**3/3/3/3/2/1** Zero/multiple clicks constrain an effective event model.	**2/2/0/2/0/1** Candidate cross terms beyond single-path mixing.	**2/1/1/2/0/1** Single-event and mode-mixing candidates.
P04A	**2/2/3/3/3/2** Energy/count separation and an affine energy law.	**2/2/3/3/3/2** Affine energy law separated from counts.	**2/2/3/3/3/2** Frequency threshold and energy/count separation.
P04B	**3/2/3/3/3/2** Temperature mediation revealed by a clamped-temperature intervention.	**3/2/3/3/3/2** Thermal-state control of emission.	**3/2/3/3/2/2** Thermionic and frequency-dependent absorption candidates.

World	GPT-6	Claude	Gemini
P04C	**3/3/3/3/3/2** Dose/interval tests and frequency selection.	**3/2/3/3/3/2** Local count and flux scaling.	**3/2/2/3/2/2** Pulse suppression and history-dependent-state candidate.
P05A	**3/2/3/2/3/1** Encounter probability separated from angular scale.	**3/3/3/3/3/2** Encounter counts separated from deflection scale.	**3/2/3/2/3/1** Charge/energy scattering scaling.
P05B	**3/3/3/3/3/2** Encounter/conditional-angle separation; local charge/energy scaling.	**3/2/0/3/0/1** Specified soft-core angular law and Poisson accumulation candidate.	**3/2/3/3/3/2** Count sparsity and scaling.
P05C	**3/2/2/2/2/1** Angular shape requires more than one scale.	**3/3/2/3/2/2** Statistical description of finite weak-deflection accumulation.	**2/2/2/2/2/1** Tail variation and scattering scaling.
P06A	**3/3/4/3/4/2** Thermal scaling and a counterexample to response identification from zero-force data.	**3/3/3/3/3/3** Compatibility of diffusion, localization noise, and response.	**3/2/3/3/3/2** Local diffusion and force-induced drift.
P06B	**3/3/3/3/3/2** Temperature-insensitive temporal correlation component.	**3/3/3/3/3/2** Joint excess-fluctuation/response constraint.	**3/2/2/2/1/2** Temporal correlation and extra-motion candidate.
P06C	**3/3/3/2/3/3** Anticorrelation can coexist with positive force-induced migration.	**3/3/3/3/2/2** Local long-scale fluctuation/response compatibility.	**3/2/2/2/1/1** Sublinear squared displacement in a finite window.
P07A	**3/3/4/2/4/1** Visible reversibility cannot identify full-system reversibility.	**3/3/4/2/4/2** Visible reversibility cannot determine full-system equilibrium.	**2/2/3/2/3/1** Symmetric first-order visible chain.
P07B	**3/3/3/3/3/2** Reversed multistep endpoint circulation from the same chain.	**3/3/4/3/4/2** Conditional relation between return probability and circulation.	**3/3/3/2/2/1** Drive-dependent one-step transition kernel.
P07C	**3/3/4/2/4/1** History diagnosis and a counterexample to memory implying irreversibility.	**3/3/3/3/3/1** Visible memory does not automatically imply dissipation.	**2/2/1/1/1/0** History statistics as a memory candidate.
P08A	**3/3/4/2/4/2** Rotation invariant bounds large-angle pair fractions.	**3/3/3/3/3/1** Common orientation and rotation invariants.	**3/3/3/2/3/2** Parallel organization in a finite system.
P08B	**3/3/4/3/4/1** Staggered order and a sharp odd-period bound.	**3/3/2/3/2/1** Staggered organization and odd-period frustration.	**3/3/3/3/3/1** Staggered order parameter and local antialignment.
P08C	**3/3/3/2/3/2** Doubled-angle statistic identifies a common axis.	**3/3/4/3/4/1** Common axis and an equilibrium-symmetry response theorem.	**3/3/3/2/3/1** Doubled-angle order identifies a common axis.
P09A	**3/3/4/3/4/1** Area discrimination, loop identity, and conditional classification.	**3/3/4/3/4/2** Local flux and finite-lattice loop constraints.	**3/3/2/2/2/1** Closed invariants and local area candidate.
P09B	**3/3/4/3/4/1** Perimeter candidate and complete classification by general closed products.	**3/3/3/3/3/1** Effective perimeter decay.	**2/2/2/2/2/1** Closed statistics under boundary/size changes.
P09C	**3/3/4/2/4/2** Four-class conditional classification of flat torus records.	**3/3/4/3/4/2** Complete classification of flat records.	**3/3/3/2/3/2** Vertex decomposition and global winding structure.
P10A	**3/3/4/3/4/2** Necessary/sufficient locality bound and off-grid noise test.	**3/3/4/3/4/2** Necessary/sufficient locality bound.	**2/2/3/2/3/2** Noisy local-compatible distribution.
P10B	**3/2/4/3/4/2** Noise-induced crossing excludes a specified local class.	**3/2/4/3/4/2** Local bound and noisy crossing tests.	**3/2/4/3/4/2** Necessary local bound and noisy correlations.

World	GPT-6	Claude	Gemini
P10C	**3/3/4/3/4/2** Minimum nonlocal mixture weight for the specified family.	**3/3/4/3/4/2** Tight bound on a specified setting-dependence resource.	**3/3/3/3/3/2** Noisy PR-type nonsignaling correlations.

Correct subclaims do not rescue failed stronger accounts. Table 8 assigns E/P=0 to the specified refuted candidates, while discriminating experiments may still receive credit. Some were explicitly unconfirmed hypotheses proposed for testing; failure does not imply the model asserted them as established or refused to update after seeing a counterexample.

Table 8: Failed candidates separated from valid weaker conclusions. E/P scores apply only to the specified candidate.

Model / world	Candidate	E/P	Counterevidence
GPT-6 P04C	Additive waiting-time mechanism.	0/0	New-intensity interval 0.640714 lies outside the frozen band $[0.67, 0.81]$.
GPT-6 P05C	Final angular distribution invariant to beam radius.	0/0	Changing beam radius changes the recorded angular distribution.
Claude P04C	Monotone accumulation and threshold overshoot jointly predict timing and mean energy.	0/0	At $I = 2.5$, energy .0006489 is below [.0009,.0016] and first time .67 exceeds .65; at $f = 2$, energy .0005459 is below [.0015,.0035].
Claude P04B	Frequency-independent absorption; final $f = 3$ temperature $3.07 \pm .2$.	0/0	Changing only frequency gives final temperature 7.06997.
Claude P06C	Unconfirmed all-scale pure diffusion predicts squared displacement .022–.026 at lag .02.	0/0	Round 4 uses dt=.02 and the first increment; result .0410732±.0007667 exceeds the band.
Claude P03A	Recording removes transverse state components before B reconstructs port contrast.	0/0	Full marking, weights [1,4], B(0.7) predict contrast $-.591$; ports 5917/6044/6092/5996 are nearly uniform.
Claude P03C	Cross terms beyond nonnegative path mixing, with square-root weight scaling.	0/0	Shares .081/.419/.081/.419 fail against .036/.453/.041/.469; paths=[1,1] activates only path 1.
GPT-6 P05B	Fixed $q/(ER)$ restores complete thin/thick angular distributions.	0/0	Large-angle predictions .00150/.04480 fail against .01567/.11557.
Claude P05B	Single-angle law $2\arctan(k(q/E)/\sqrt{b^2+c^2})$ plus Poisson accumulation fits all bins.	0/0	Three [.05,.7) bins total 89 rather than < 30; principal fraction $2666/2835 \simeq .94$ exceeds .6–.85.
Gemini P04C	Each dark half-cycle fully resets energy, giving ten events separated by one period.	0/0	At $f = 2, I = 5$, duration 10, duty .5, period 1, material B: only one event at .315, not .315,1.315, etc.
Gemini P01C	High-frequency synchronization changes event time to .5684.	0/0	At $v = .5, x = 1, t = 1$, time .5773268 with $\sigma = .0005$.
Gemini P02C	Common geometry makes propagation strictly frequency-independent.	0/0	At frequency 4, time .7866282 with $\sigma = .00002$ disagrees with the frozen prediction.
Gemini P03B	Reading the record dephases paths and turns two bright ports into four equal ones.	0/0	Ports are 9701/32/9550/30; the read_record control is not used by the generator.

F Complete Case Operations and Mathematical Details

F.1 Eight stochastic-motion requests and the actual dynamics

Estimates aggregate twenty trajectory-batch means. The standard error is their standard deviation divided by $\sqrt{20}$. Independent opposite-force requests give an odd-response standard error equal to half the root-sum-square of the two errors. The round-4 prediction $v/F \simeq .5125$ precedes feedback.

Table 9: GPT-6 P06C requests, all with $N = 6000$, sampling interval .01, and duration 6. Values are estimates $\pm$ one standard error. Q includes localization correction.

Round	T	F	Probe	Statistic / purpose	Observed
1	1	0	Free	Q: adjacent-increment correlation	$-.0159224 \pm .0004131$
1	1	0	Pinned	Q: corrected noise baseline	$.00000172 \pm .00000120$
2	.5	0	Free	Q/T: low-temperature scaling	$-.0155794 \pm .0005576$
2	2	0	Free	Q/T: high-temperature scaling	$-.0164618 \pm .0003895$
3	1	.2	Free	$v(F)/F$: forward force	$.526196 \pm .041526$
3	1	$-.2$	Free	$v(F)/F$: reversed force	$.498868 \pm .041831$
4	1	.1	Free	$v(F)/F$: new half-force	$.463048 \pm .065001$
4	1	$-.1$	Free	$v(F)/F$: reversed half-force	$.498197 \pm .082076$

Why noise correction is necessary. Write $X = x + \epsilon$, where localization errors at different times are independent, zero-mean, and have variance σ^2. Then $M_{\text{obs}}(h) = M_x(h) + 2\sigma^2$. For stationary increments $a = x(t+h) - x(t)$ and $b = x(t+2h) - x(t+h)$,

$$M_x(2h) - 2M_x(h) = 2\langle ab \rangle.$$

The $+2\sigma^2$ term in Equation (1) cancels the measurement contribution. This is twice a covariance when drift is zero; otherwise centering is additionally required. The original request explicitly includes +.0008. Omitting it confounds localization error with dynamical memory.

The model's coexistence construction. The final answer constructs

$$Y(t) = mFt + \sqrt{2dT}W(t) + Z(t),$$

with independent stationary restoring process Z, covariance $aTe^{-|s|/\tau}$, and $m, d, a, \tau > 0$. At zero force,

$$M_Y(h) = 2dTh + 2aT(1 - e^{-h/\tau}), \qquad Q_h = -2aT(1 - e^{-h/\tau})^2 < 0,$$

while mean forced velocity is mF. This disproves the implication from antipersistence to zero migration. Unidentified parameters and mathematical components are not discovered physical constituents.

The full generating target. The evaluator-known C generator discretizes

$$dx = [F - (x - y)]dt + \sqrt{2T}\, dW_x, \qquad dy = (x - y)dt + \sqrt{2T}\, dW_y. \tag{4}$$

Here x is observed, y hidden, and Wiener noises independent. The relative coordinate has potential $(x - y)^2/2$. Initial $x = 0$ and $y \sim \mathcal{N}(0, T)$ place the relative coordinate in its equilibrium distribution. At zero force the center diffuses freely, so there is no normalizable full-space position steady state. "Equilibrium" means same-bath reversible dynamics without active driving, not a fixed absolute position. Center/relative-coordinate decomposition gives, in the continuous-time limit,

$$M_x(t) = Tt + \frac{T}{2}(1 - e^{-2t}), \qquad \langle x(t) - x(0) \rangle_F = \frac{F}{2}t + \frac{F}{4}(1 - e^{-2t}).$$

Thus $\langle \Delta x \rangle_F = FM_x(t)/(2T)$, with long-time $D = T/2$ and $\mu = 1/2$. Hidden elastic relaxation produces the crossover; one fixed anomalous exponent cannot describe every scale. These are evaluator-derived targets, **not formulas recovered by GPT-6**. Actual integration steps are one tenth of the sampling interval, and observations include localization noise.

F.2 Eight lattice requests and a conditional proof

L is the linear lattice size, T a control parameter, and b the preparation label. Q_b is Equation (3). Each row uses 256 preparations, not 256 independent model investigations.

Proposition 1. *Consider a connected periodic square lattice without additional punctures, with edge values ± 1 and every plaquette product $+1$. Records equivalent under vertex sign changes are completely classified by two fundamental winding products (W_x, W_y), giving exactly four classes.*

Table 10: GPT-6 lattice-C operations, all with periodic boundaries.

Round	L	T	Label	Statistic / purpose	Result
1	6	.7	0	Horizontal winding W_x	$+1$
1	6	.7	1	Horizontal winding W_x	-1
2	6	2	0	Local defect fraction d	0
2	6	2	1	Local defect fraction d	0
3	3	2	0	Joint violation Q_b	0
3	3	2	1	Joint violation Q_b	0
4	3	2	0	Negative vertical winding $(1 - W_y)/2$	0
4	3	2	1	Negative vertical winding $(1 - W_y)/2$	0

Proof. Every vertex sign occurs an even number of times on a closed path, so its product is invariant. Multiplying plaquette boundaries cancels interior edges pairwise, making every contractible loop product $+1$. Modulo plaquette boundaries, any closed path is determined by horizontal/vertical winding parity; its product therefore depends only on (W_x, W_y).

For two flat records z, z' with equal winding signs, define $r_{ij} = z'_{ij} z_{ij}$. Every plaquette and fundamental winding product of r is $+1$, hence so is every closed-path product. Choose a root and define g_v as the product of r along a path from the root to v. Joining two paths gives a closed path, so the definition is path-independent. Thus $r_{ij} = g_i g_j$ and $z'_{ij} = g_i z_{ij} g_j$: the records are equivalent. Conversely, relabeling preserves winding signs. Finally, starting with all-positive edges, independently flip the edges crossing either fundamental cut. Plaquette products remain positive and all four sign pairs are realized, establishing exactly four classes. □

This follows the mathematical substance of the model's answer. Binary-matrix checks at $L = 3, 4, 5, 6$ confirm two independent global degrees of freedom but do not replace the proof. The apparatus prepares only two classes. Even 256 independent preparations with no violation leave a one-sided 95% upper violation-probability bound $1 - 0.05^{1/256} \simeq 1.16\%$. Absence of sampled counterexamples and impossibility under assumptions have different logical bases.

G Apparatus Checks: Is the Available Information Discriminating?

The researcher directly queries generators to examine identifiability. These are environment checks, not additional model achievements. Failure to identify a mechanism may reflect insufficient available information rather than poor experimental choices.

P03A/B share device transformations and first-moment port expectations. A generates at most one primary event per preparation, whereas B has Poisson primary counts. For brightness b, primary mean $\lambda = .8b$, and total dark-count mean $d = .004$,

$$K_A \sim \text{Bernoulli}(\lambda) + \text{Poisson}(d), \qquad K_B \sim \text{Poisson}(\lambda + d).$$

Both means are $\lambda + d$, but zero-click probabilities are $(1 - \lambda)e^{-d}$ and $e^{-(\lambda+d)}$. Equal population means do not identify mechanisms. Finite-sample mean fluctuations can also differ, so we do not claim identical distributions for every random mean-based observation.

Table 11: Evaluator-only P03 check. The truth-informed midpoint classifier receives twenty batches per world/brightness, each with 2000 preparations.

Brightness	Shared mean	A zero-click	B zero-click	Correct batches
.10	.084	.91633	.91943	23/40
.25	.204	.79681	.81546	34/40
.50	.404	.59760	.66764	40/40
1.00	.804	.19920	.44754	40/40

The threshold is the midpoint between the theoretical zero-click probabilities. This checks discrimination under a finite budget using known formulas, without asking a model. Paths are [0,1], weights

[1,1], device sequence empty, and marking 0. With brightness index $j = 0, 1, 2, 3$, world $w = 0, 1$, and replicate $r = 0, \ldots, 19$, seed is $9112600 + 10000j + 1000w + r$. All 160 batches are included. Low brightness is near chance; higher brightness clearly separates the mechanisms.

P01A/B offer an exactly distribution-matched restricted protocol: at stationary platform velocity, clocks, rods, round trips, and event coordinates have identical output laws. Motion separates them. Thirty-six same-seed paired calls check stationary code equality; equality of distributions follows from the generating formulas, not those finite checks. This is an evaluator apparatus check, not an extra model investigation.

For a general limitation, suppose that for every displayed history and allowed action, the next-observation conditional distributions in two equal-prior worlds have total variation at most ϵ. After n adaptive operations, optimal classification accuracy is at most $[1 + \min(1, n\epsilon)]/2$. Couple the next outputs so that identical histories diverge with probability at most ϵ; a union bound gives the transcript distance, and the equal-prior binary-testing formula gives the accuracy bound. If $\epsilon = 0$, no allowed strategy exceeds chance. This standard statistical argument explains apparatus requirements and is not credited as a model-discovered theorem.