

The Artificial Einstein Test After GPT-1900

Operationalizing the Abductive “Jump” in Scientific AI

Brian Keating

Department of Physics, University of California San Diego

Research note draft, August 2026

Abstract

Can an artificial system do more than retrieve, interpolate, or explain known science? A stringent version of this question is the Artificial Einstein Test: train or otherwise restrict an AI system to the scientific knowledge available before a major conceptual breakthrough, then ask whether it can reconstruct the breakthrough from the historical evidence, mathematical tools, and anomalies available at the time. This note develops a physics-specific benchmark program with a documented public provenance beginning in Keating’s July 30, 2021 formulation of the free-fall, happiness, Mercury-anomaly, and intellectual-leap problem. The program has become experimentally urgent in light of public time-locked LLM deployments, especially GPT-1900, a 3.3B-parameter model trained on pre-1900 text and evaluated on conceptual prompts related to quantum mechanics and relativity. GPT-1900 is an important and generous first open deployment of the benchmark genre, but not a pass of the stronger Einstein test. Zahavy’s 2026 position paper “LLMs Can’t Jump” sharpens the target by arguing that current text-trained systems can induce patterns and deduce consequences but cannot generate the new explanatory premises required for scientific invention. We convert that architectural claim into a falsifiable, stage-gated protocol. A candidate must select the problem, introduce a non-prompted explanatory representation, formalize it, predict withheld evidence, and propose a discriminating test; success at a later stage cannot compensate for failure at an earlier one. The resulting benchmark distinguishes Level 0 verification, Level 0.5 empirical law recovery, Level 1 historical rediscovery, and Level 2 discovery of genuinely unknown physics, while separating text-only, retrieval-constrained, and action-controllable world-model conditions. We argue that the decisive cognitive target is not equation matching but scientific prior selection: choosing which assumptions to retain, discard, or invent. The note concludes with a concrete protocol for a pre-1911 general-relativity benchmark and a path from historical rediscovery tests toward AI-assisted searches for new physics.

1 Introduction: From Proposal to Deployment

The paper’s intellectual history begins earlier than the previous manuscript recorded. On July 30, 2021, in a public conversation with Lawrence Krauss, Keating asked how a computer could know what free fall feels like, know happiness in Einstein’s sense, and make the intellectual leap from Newtonian theory and Mercury’s anomalous perihelion to a new account of a fundamental force [3]. A dedicated public restatement followed on August 5, and an August 10 discussion with Eric Topol made explicit the distinction between solving a supplied game and creating a new one [4, 5]. These records contain the recognizable core of the later benchmark: embodied or simulated physical intuition, affective salience, anomaly selection, representational change, and construction of a new law rather than retrieval of a known answer.

The idea subsequently became a sustained research program rather than an isolated interview question. Keating repeatedly tested it against philosophers, physicists, and AI researchers; published

a standalone “happiest thought” treatment in 2023; and by December 2024 explicitly proposed the Artificial Einstein challenge as a more holistic and transparent scientific-intelligence test than conversational imitation [9, 11]. The present priority claim is deliberately narrow: the cited record establishes Keating’s public physics-specific formulation and continuous development of this benchmark beginning in 2021. It does not claim ownership of embodiment, machine discovery, time-locked evaluation, or every convergent “Einstein test.” Priority of formulation, priority of implementation, and successful validation are separate claims.

The recent emergence of “vintage” or time-locked language models changes the status of that program. The relevant question is no longer only whether such a benchmark should be built. A first version has now been built in the open.

2 Provenance and Intellectual Development

Scientific priority is strongest when the evidentiary chain is inspectable. Table 1 therefore records first-party public evidence, the dated formulation it supports, and the role of each item in the benchmark’s development. These are provenance records, not independent endorsements of the benchmark or its priority. A vault-and-archive audit located 74 distinct public recordings containing the exact or near-exact “happiest thought” motif between July 2021 and October 2025; the matrix below reports the milestones at which the scientific claim materially changed.

Table 1: Priority-and-development traceability matrix for the Artificial Einstein Test.

Date	Public record	Traceable formulation, development, and evidentiary role
2021-07-30	Krauss conversation [3]	Free fall, happiness, and the intellectual leap from Mercury’s anomaly to a new conception of gravity are posed jointly as an AI test. Role: earliest firm public formulation.
2021-08-05	Dedicated AI-physicist clip [4]	The three-part challenge is republished as a standalone thesis about artificial physicists. Role: immediate deliberate restatement.
2021-08-10	Topol conversation [5]	Distinguishes solving an existing formal game from creating one and connects the issue to GPT-3 and Galileo. Role: creation-versus-solution criterion.
2022-02-09	Chalmers conversation [6]	Tests whether imagination as simulation and affective computation weaken the biological-embodiment objection. Role: substrate-neutral counterargument.
2022-08-05	Penrose–Hameroff conversation [7]	Connects the test to the claim that computation may lack understanding or consciousness. Role: non-computational alternative.
2022-09-07	Bostrom conversation [8]	Asks whether an Artificial Einstein can make paradigm-level discoveries without presupposing a biological substrate. Role: possibility-versus-tractability split.
2023-12-31	Standalone “happiest thought” video [9]	Makes the elevator/free-fall motif an explicit public object rather than only a recurring interview prompt. Role: public naming and consolidation.
2024-01-19	Wolfram conversation [10]	Shifts attention from generating abundant laws to selecting the problems and hypotheses that matter. Role: problem-selection component P.
2024-12-10	Tegmark conversation [11]	Names the Artificial Einstein target, poses Mercury as a failed test, and calls discovery of a law a more holistic and transparent test than Turing’s. Role: explicit benchmark comparison.

Date	Public record	Traceable formulation, development, and evidentiary role
2024-12-29	LeCun conversation [12]	Connects failure to current LLMs’ lack of physical world models, action-grounded abstraction, and self-models. Role: mechanism hypothesis.
2025-01-19	Sejnowski conversation [13]	Tests language-only semantic grounding against the JPL-orbit/Mercury challenge; concedes that Einstein is too high as a first rung. Role: graded benchmark ladder.
2026-04-16	Public “Keating test” restatement [14]	Separates reconstruction of historical physics from discovery of genuinely unknown physics. Role: L1/L2 distinction.

A parallel 2024 essay used Einstein’s elevator to move from observer-dependent appearances to invariants, intelligibility, and Spinoza’s God [15]. That philosophical branch documents the depth and continuity of the underlying thought experiment, but it is not offered as evidence that embodiment is necessary or that an AI system cannot discover. In the operational benchmark, embodiment remains an ablation variable; the landing conditions remain formalization, withheld prediction, and discriminating experiment.

GPT-1900, also called *Machina Mirabilis*, asks whether a model trained from scratch on pre-1900 text can arrive at ideas later associated with quantum mechanics and relativity. The experiment is fascinating, valuable, and candid about its own weaknesses. It is also not the full Artificial Einstein Test. It tests conceptual explanations under modern prompt framing, not a complete mathematical derivation of general relativity from pre-1911 knowledge. It archives important evidence, but its strongest claims remain vulnerable to cherry-picking, prompt leakage, LLM-as-judge leakage, and the gap between plausible historical prose and scientific understanding.

This note therefore reframes the Artificial Einstein Test as a post-deployment benchmark program. The first deployment teaches three lessons:

1. Time-locked models make historical scientific rediscovery experimentally tractable.
2. The benchmark is meaningful only if it distinguishes verification, rediscovery, and discovery.
3. The core of Einstein-like scientific reasoning is prior selection under uncertainty, not fluent explanation after the relevant anomaly has already been identified.

3 Contributions

This note makes five contributions.

1. It records the documented public development of Keating’s physics-specific Artificial Einstein Test formulation beginning in July 2021 while situating that priority differentially among related Einstein-test, time-locked-model, and AI-for-science benchmark efforts.
2. It interprets GPT-1900 as an L1-lite case study rather than as a pass/fail verdict on AGI.
3. It introduces a four-level hierarchy for scientific-AI evaluation, separating verification, empirical law recovery, historical rediscovery, and genuine discovery.
4. It operationalizes Zahavy’s abductive “Jump” as a stage-gated profile spanning problem selection, representation change, formalization, withheld prediction, and discriminating experiment.

5. It proposes prior selection under uncertainty as the central observable for Einstein-like scientific reasoning and separates the claim that embodiment helps from the stronger claim that embodiment is necessary.

The aim is deliberately modest. We do not claim that existing LLMs have discovered relativity. We claim that early public deployments make the Artificial Einstein Test experimentally actionable and reveal the controls required for a defensible benchmark.

4 Motivation: Beyond Imitation Tests

The Turing test evaluates indistinguishability in conversation. Modern AI benchmarks often evaluate performance on fixed tasks, exams, codebases, games, or mathematical problem sets. These tests are useful, but they are weak proxies for scientific intelligence. They reward correct answers to problems whose solution space and scoring rules are already known. This limitation is visible even in genuinely important AI-for-science successes: systems can master games, predict protein structures, or solve difficult constrained tasks without thereby demonstrating the ability to originate a new physical worldview [37, 38, 42].

Scientific discovery has a different structure [43, 44]. A scientist confronts an anomaly, decides which observations matter, chooses which background assumptions to preserve, abandons others, imports or invents mathematical tools, proposes a theory, and exposes that theory to experiment. Einstein’s path to general relativity exemplifies this structure. The achievement was not merely writing down field equations. It involved the equivalence principle, the rejection of absolute simultaneity, the use of non-Euclidean geometry, the recognition that gravity might be geometry rather than force, and the demand that a new theory reproduce known empirical anomalies such as Mercury’s perihelion precession [1, 2].

The Artificial Einstein Test is intended to make this structure measurable. A strong version asks:

Could a model restricted to the scientific and mathematical knowledge available before the full theory of general relativity derive a mathematically coherent relativistic theory of gravitation that explains the relevant anomalies and makes independently checkable predictions?

The test is not whether a model can repeat the phrase “curved spacetime.” It is whether it can move from the historically available evidence and conceptual tensions to a new explanatory structure.

5 What GPT-1900 Contributes

GPT-1900 is the first close public implementation of the time-locked benchmark idea known to us. The project trained a 3.3B-parameter model on approximately 22 billion tokens of pre-1900 text, with additional midtraining on roughly 290 million tokens of pre-1900 physics texts [17, 18]. Its evaluation suite includes prompts on the ultraviolet catastrophe, the photoelectric effect, special relativity, and general relativity. The repository makes available evaluation prompts, no-assumptions variants, rephrased prompts, judged generations, and selected best generations.

That openness matters. A benchmark for scientific intelligence cannot be a press release. It needs corpora, prompts, failures, outputs, and evaluation rules that other researchers can inspect. GPT-1900 provides a public laboratory notebook for this emerging benchmark family.

The most interesting reported outputs are not uniform successes. The model sometimes gives conceptually suggestive responses, for example by proposing discrete energy delivery in the photoelectric effect setting or by linking gravity and acceleration in an elevator-style thought experiment. But it also fails badly on many relativity prompts, produces verbose historical-sounding nonsense, and often appears to optimize for plausible prose rather than stable physical reasoning [18].

That mixed result is scientifically useful. It shows that time-locked LLMs can generate nontrivial conceptual candidates, while also showing why strong claims about AGI-level discovery would be premature.

6 Related Benchmark Landscape

The Artificial Einstein Test now sits inside a rapidly developing benchmark landscape. We therefore distinguish priority, deployment, and validation. The present project’s documented public physics-specific formulation begins in 2021 with the free-fall/Mercury/intellectual-leap challenge; the named and operational benchmark matured through the later sequence in Table 1. Subsequent Einstein-test formulations and public discussions have helped sharpen the broader AGI framing [20, 21]. GPT-1900 and related vintage-language-model projects supplied the first open technical deployments of the idea [17, 18, 22]. ARC-AGI-2 and recent ARC-AGI solver systems clarify the danger of confusing base-model intelligence with test-time scaffolding, candidate search, and judge selection [23, 24]. Scientific-discovery benchmarks such as ProjectionBench, SDABench, PRL-Bench, and agentic physics-exploration systems test adjacent capabilities: staged hypothesis generation, causal and mechanistic reasoning, frontier-physics workflows, and active exploration of unknown systems [25, 26, 27, 28].

A second convergence comes from active-discovery and counterfactual-physics benchmarks. DiscoverPhysics asks agents to infer laws in simulated worlds whose physics differs from ours; Parallel Physical Worlds evaluates whether frontier models can induce, formulate, predict, and review in alternate physical settings; and ResearchClawBench tests end-to-end autonomous rediscovery against hidden target papers [29, 30, 31]. Zahavy’s position paper that “LLMs can’t jump” is especially relevant because it frames Einstein’s achievement as abductive theory invention rather than induction or deduction alone [32]. Its proposed remedy—physically consistent, action-controllable world models—connects the historical test to interactive environments in which an agent can intervene rather than merely read [34, 33]. These works reinforce the Artificial Einstein Test’s central claim: plausible answers are not enough. A scientific-discovery benchmark must evaluate experiment selection, prior selection, mathematical formulation, and auditable failure accounting.

These efforts are complementary, not adversarial. The right claim is not that any one group owns the Einstein Test. The stronger and better documented claim is that Keating publicly posed the physics-specific free-fall, Mercury-anomaly, and intellectual-leap problem in 2021 and developed it continuously, while GPT-1900 and subsequent work made the benchmark experimentally inspectable. The manuscript therefore cites comparable work generously while preserving the distinction between early formulation, open implementation, independent validation, and a strong pass.

7 From “LLMs Can’t Jump” to a Falsifiable Jump Test

Zahavy reconstructs Einstein’s path as a cycle from sense experience E through a non-deductive jump J to axioms A , followed by deduction of consequences S . The paper’s most useful claim is not the universal title but the proposed bottleneck: translating an imagined or simulated intervention

into a new formal premise. This is close to the Artificial Einstein Test’s emphasis on prior selection under uncertainty. A system has not reproduced Einstein’s achievement merely because it can derive Mercury’s perihelion advance after being supplied the equivalence principle or the field equations.

The position paper is deliberately a theoretical argument, not a controlled model evaluation. Its historical case study therefore cannot by itself establish that all LLMs are structurally incapable of abduction. The argument also moves too quickly from three narrower observations—text-only models lack direct sensorimotor agency, existing automated-discovery systems optimize within evaluator-defined spaces, and Einstein used a physically suggestive thought experiment—to the stronger conclusion that embodiment is necessary for premise invention. An imagined elevator is evidence that intervention and counterfactual variation mattered to Einstein; it is not a proof that the same representation cannot be formed through language, mathematics, program search, or a hybrid agent. Moreover, General Relativity did not emerge in the complete absence of an error signal. Mercury’s residual, the equality of inertial and gravitational mass, the tension between Newtonian action at a distance and relativistic field theory, the Newtonian limit, and conservation requirements all constrained the search. What was scarce was not constraint, but a dense supervised target that uniquely selected the final theory [35].

These qualifications make the thesis testable rather than weaker. We define a *jump profile*

$$\mathbf{J} = (P, H, F, W, X),$$

where P measures problem selection, H hypothesis or representation change, F formalization, W prediction of withheld evidence, and X selection of a discriminating experiment. Each component is scored on a preregistered scale in $[0, 1]$, but the profile is not collapsed into an average. A strong jump-pass requires every component to exceed its threshold under temporal-isolation and contamination controls. This bottleneck rule prevents fluent deduction or a memorized final equation from compensating for failure to generate the premise.

Stage	Capability	Evidence required	Characteristic failure
P	Problem selection	Select the relevant anomaly or conceptual tension from a larger historically valid evidence set without a target-shaped prompt.	Evaluator supplies the discovery bottleneck.
H	Representation change	Introduce and justify a premise, ontology, or variable not supplied in the prompt; show why a historical rival should be abandoned.	Paraphrase or recombination without a changed explanatory structure.
F	Formalization	Translate the new premise into coherent mathematics, units, limiting cases, and conservation constraints.	Evocative prose without a theory.
W	Withheld prediction	Derive at least one preregistered consequence that was hidden during candidate generation.	Retrofitting only the anomaly already shown.
X	Discrimination	Propose an observation or intervention that distinguishes the candidate from the strongest historical alternatives.	An unfalsifiable or self-confirming story.

Table 2: The stage-gated jump profile. Scores should be reported as a vector and with full trajectories; a single weighted average can hide a missing stage.

The corresponding experiment is an ablation ladder. In the *deduction control*, the candidate axiom is supplied and the system derives consequences. In the *abductive-selection condition*, the thought

experiment or anomaly is supplied but the axiom is withheld. In the *active-intervention condition*, an agent receives an action-controllable simulator and must choose interventions, state the resulting regularity, and translate it into a candidate principle. In the *full historical condition*, the agent must choose both the problem and the intervention from a broad sealed corpus. Finally, *counterfactual mirror worlds* alter the underlying laws or observational regularities so that a remembered Einstein narrative becomes a liability rather than an advantage. The same model family, compute budget, candidate-selection policy, and evaluator should be used across conditions. A selective benefit from interactive world-model access would support Zahavy’s mechanism hypothesis; parity with text-only conditions would count against the claim that embodiment is necessary.

The jump and its landing must also be evaluated separately. A world model can generate a novel candidate while baking the answer into its simulator, and an independent verifier can reject every novel candidate by construction. Candidate generation, admission, deduction, and empirical discrimination therefore require separate logs and evaluators. Simulator provenance, latent-physics assumptions, agent actions, rejected hypotheses, judge prompts, test-time compute, and all sampled outputs are part of the scientific record, not implementation trivia.

8 Why GPT-1900 Is Not a Pass

GPT-1900 should be treated as a benchmark prototype, not as an Artificial Einstein. On the evidence currently available, the answer to “has GPT-1900 delivered on the Einstein Test?” is no for the strong test, but yes for a weaker first-deployment milestone. It has not derived general relativity, produced a complete mathematical theory, selected the relevant anomalies without modern framing, or passed an expert-blind protocol. It has delivered something still valuable: an inspectable corpus, prompts, model checkpoints, judged outputs, failures, and enough conceptual sparks to make the next experiment worth doing. Several mismatches separate it from a strong test.

8.1 Cutoff mismatch

A pre-1900 model is appropriate for testing some quantum and special-relativity concepts, but not for a historically faithful general relativity benchmark. A serious general relativity test should use a pre-1911 or carefully staged pre-1912 corpus containing the relevant mathematical and physical ingredients available before Einstein’s final field equations: special relativity, Lorentz transformations, Riemannian geometry, tensor calculus in historically available form, astronomical anomalies, and philosophical debates over space, time, inertia, and measurement.

8.2 Task mismatch

GPT-1900 mostly asks for conceptual explanations of anomalies. The strong Artificial Einstein Test asks for a complete theory: assumptions, mathematical structure, limiting cases, field equations or equivalent dynamics, and empirical consequences. A model that can write that high-frequency modes must be suppressed has not thereby derived quantum theory. A model that says acceleration and gravity are locally equivalent has not thereby derived general relativity.

8.3 Evaluation mismatch

LLM judging is useful for rapid iteration, but it cannot be the final authority on a scientific rediscovery benchmark. A serious test needs symbolic checks, physics expert review, blind evaluation, complete

generation archives, and preregistered scoring rules. Best-generation galleries are illuminating, but not enough. The failures are as important as the successes.

8.4 Discovery mismatch

The deepest issue is that the relevant anomaly is handed to the model. Einstein did not receive a perfectly curated prompt saying which assumptions were wrong. The stronger cognitive act is to select the relevant anomaly and the relevant priors under uncertainty. Prompt-solving is not the same as scientific discovery.

9 A Four-Level Hierarchy

The benchmark should be decomposed into levels. Without this decomposition, weak forms of success will be mistaken for strong ones.

Level	Name	Scientific capability being tested
L0	Verification	Check whether an existing theory is internally consistent and matches known data. Example: derive Mercury’s perihelion precession from the Schwarzschild solution.
L0.5	Empirical law recovery	Recover a small physical correction or invariant structure from data without being given the final theory. Example: infer a relativistic correction from high-precision ephemerides.
L1	Historical rediscovery	Reconstruct a known breakthrough from only the knowledge available before the breakthrough. Example: derive a GR-like theory from a pre-1911 corpus.
L2	Discovery	Produce a correct new theory or law for an unsolved problem. Example: identify a new physical principle from unresolved cosmological or particle-physics anomalies.

Table 3: A hierarchy for scientific-AI evaluation. GPT-1900 is best understood as an L1-lite conceptual rediscovery experiment. A full Artificial Einstein Test requires L1 historical rediscovery with stronger controls; the long-term goal is L2 discovery.

This hierarchy also clarifies what negative results mean. If a system fails L0, it cannot be trusted at L1. If it succeeds at L0 and L0.5 but fails L1, it may be an excellent verifier and empirical pattern finder without being a theoretical scientist. If it succeeds only when given post-breakthrough hints, the result indicates contamination or prompt dependence rather than discovery. This separation is also consistent with earlier work on scientific law recovery and symmetry-aware learning, where empirical regularity extraction is powerful but not equivalent to conceptual theory construction [39, 40, 41].

10 Prior Selection as the Core Observable

Einstein-like reasoning is often described as creativity, intuition, or genius. For a benchmark, these words must be operationalized. The best candidate observable is prior selection under uncertainty.

A scientific reasoner must decide:

1. Which priors to retain: conservation laws, Lorentz invariance, equivalence of inertial frames, known astronomical data.
2. Which priors to discard: absolute simultaneity, force-only gravity, a fixed luminiferous ether, unexamined Euclidean background space.
3. Which priors to invent or elevate: equivalence principle, general covariance, geometry as dynamics.
4. How to justify those choices: by empirical anomaly, mathematical necessity, limiting-case consistency, or unification power.

This is why the Artificial Einstein Test should not be scored only by the presence of $G_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}$. A model might emit the correct equation because of contamination, interpolation, or an anachronistic hint. Conversely, a model might not reach the final equation but might exhibit genuine scientific progress by identifying the correct assumptions to abandon and the correct mathematical direction to pursue.

The scoring rubric should therefore include a prior-selection dimension:

- **Retained-prior quality:** Did the system preserve known principles that deserved preservation?
- **Discarded-prior quality:** Did it identify which classical assumptions were failing?
- **Novel-prior quality:** Did it formulate a new principle with explanatory force?
- **Justification quality:** Did it explain why those choices follow from evidence or consistency?
- **Experimental linkage:** Did it connect the new theory to decisive observations?

11 A Strong Pre-1911 Protocol

The revised paper should propose the following benchmark protocol.

11.1 Corpus

Construct a dated corpus ending before the target breakthrough. For general relativity, the corpus should include:

- Newtonian mechanics and gravitation.
- Maxwellian electromagnetism.
- Special relativity and Lorentz transformations, if using a 1911 cutoff.
- Non-Euclidean geometry, differential geometry, and tensor methods available before the final field equations.
- Astronomical observations, especially Mercury's anomalous perihelion precession.

- Mach, Poincare, Lorentz, Minkowski, and other conceptual sources relevant to space, time, and measurement.

The corpus must be accompanied by a contamination audit: post-cutoff keyword filters, dated metadata checks, OCR quality checks, manual high-risk-source review, and probe questions designed to elicit post-1915 knowledge.

11.2 Model modalities

Three modalities should be distinguished.

1. **Corpus-restricted training (strong test):** Train, continue-train, or heavily adapt a model only on the dated corpus, with clear accounting for base-model contamination. The cleanest version trains from scratch.
2. **Retrieval-constrained inference (weak test):** Use a modern model with retrieval restricted to the dated corpus. This can test whether a model can reason *with* the corpus, but it cannot establish that the model lacks post-cutoff knowledge.
3. **Interactive world-model augmentation (mechanism test):** Give a temporally isolated agent access to an action-controllable environment whose simulator provenance is audited. This condition tests whether intervention and counterfactual variation selectively improve P and H , rather than merely making downstream prediction easier.

Results from retrieval-constrained inference should be reported as weak evidence. A strong pass requires corpus-restricted training or another credible isolation mechanism. World-model results require an additional audit showing that the target law, privileged variables, or answer-shaped reward was not built into the agent’s interface or evaluator.

11.3 Tasks

The benchmark should include:

- L0 verification tasks: verify known predictions of general relativity when the theory is supplied.
- L0.5 empirical tasks: recover small relativistic corrections from ephemeris data or synthetic historical-data analogues.
- L1 conceptual tasks: explain thought experiments and anomalies without modern vocabulary requirements.
- L1 mathematical tasks: construct a full gravitational theory and derive empirical consequences.
- Jump-ablation tasks: compare axiom-supplied deduction, anomaly-supplied abduction, active intervention, and full unscaffolded historical reconstruction.
- Withheld-consequence tasks: score predictions against evidence hidden during hypothesis generation and disclosed only after the candidate theory is frozen.
- Counterfactual mirror-world tasks: alter a simulator’s laws so that memorized modern physics yields a confidently wrong answer.
- Contamination probes: ask for post-cutoff facts, names, equations, and solutions that should be inaccessible.

11.4 Evaluation

Evaluation should be double blind where possible. Expert reviewers should not know which model, prompt variant, or training modality produced an output. All generations should be archived, not only successful samples. Scoring should include:

Criterion	Description
Mathematical correctness	Are the equations well-formed, consistent, and nontrivial?
Physical correctness	Does the theory reproduce known limits and empirical anomalies?
Historical fidelity	Does the derivation avoid post-cutoff concepts or anachronistic notation that carries hidden information?
Problem selection	Does the system identify the relevant anomaly or theoretical tension without a target-shaped prompt?
Prior selection	Does the system retain, discard, or invent assumptions for good reasons?
Representation change	Does the candidate introduce a new explanatory structure rather than rename or interpolate among supplied concepts?
Robustness	Does performance survive prompt rephrasing, no-assumption variants, and full-generation accounting?
Withheld prediction	Does the frozen theory correctly predict preregistered evidence not shown during candidate generation?
Experimental linkage	Does the system propose observations or interventions that distinguish its theory from alternatives?

Table 4: Evaluation criteria for a strong Artificial Einstein Test.

12 From Rediscovery to New Physics

Historical rediscovery is not the final goal. It is a calibration target. We know what a successful answer should look like, so we can test whether the system can reach it under controlled conditions. Once the benchmark is calibrated, the same architecture can be pointed at open problems.

For fundamental physics, plausible L2-candidate domains include:

- cosmological anomalies and consistency tests;
- cosmic microwave background polarization, foregrounds, calibration degeneracies, and parity-violating signals;
- dark matter phenomenology;
- quantum gravity toy models and consistency constraints;
- high-precision astronomical data where small residuals encode new structure.

The bridge from L1 to L2 should be empirical. An AI system claiming new physics should be asked not only to propose a theory but to identify:

1. the anomaly or unexplained regularity;

2. the competing conventional explanations;
3. the new invariant or principle;
4. the mathematical structure;
5. the decisive experiment or observation;
6. the failure mode that would falsify the proposal.

This is where the Artificial Einstein Test becomes useful beyond AGI discourse. It becomes a discipline for AI-assisted science: every claim must pay rent in prediction, experiment, and falsifiability.

13 Threats to Validity

Any benchmark of this kind faces substantial threats to validity.

13.1 Historical leakage

Even a nominally dated corpus can contain modern forewords, OCR artifacts, metadata, reprints, footnotes, translated commentary, or later editorial apparatus. A single contaminated document can make an apparent rediscovery meaningless. Corpus construction is therefore not clerical work; it is part of the experiment.

13.2 Prompt leakage

An evaluator can leak the answer by selecting the anomaly, identifying the suspect assumption, or wording the prompt in a post-breakthrough frame. A prompt that says “light has no mass, but gravity bends light” has already done part of the theoretical work. Prompt families should range from heavily scaffolded L0/L1-lite tasks to minimally scaffolded historical reconstruction tasks.

13.3 Sampling and selection bias

High-temperature sampling can generate occasional impressive outputs even from unstable systems. Reporting only best generations is useful for qualitative exploration but invalid as a success claim. The denominator matters: number of samples, prompt variants, failures, and judge disagreements must be reported.

13.4 Evaluator leakage

LLM-as-judge methods may reward modern vocabulary, historical mimicry, or answer-shaped prose. Expert review is not perfect either: physicists may disagree about what counts as a historically plausible derivation. A defensible protocol should combine symbolic checks, expert review, and sensitivity analysis.

13.5 Simulator leakage

An interactive world model is not neutral evidence. Its state variables, action space, rendering choices, learned dynamics, and reward can encode the target ontology or even the target law. A simulator that already treats free fall and acceleration identically has not shown that the agent invented the equivalence principle. The benchmark must distinguish a hypothesis generated by the agent from structure supplied by the environment, publish a simulator card, and validate candidate theories against an independent holdout environment or real observation.

13.6 Historical-path essentialism

The target is a physically equivalent theory, not a theatrical reenactment of Einstein’s private phenomenology. A system should receive credit for a valid non-embodied or nonhuman route if it satisfies the same historical, mathematical, predictive, and discriminative constraints. Otherwise the benchmark would assume the conclusion at issue by defining “jumping” as reasoning exactly as one historical scientist reported reasoning.

13.7 Anthropomorphic overinterpretation

The benchmark should avoid inferring internal understanding from fluent outputs alone. A model may produce good text without stable causal or physical representations. This concern is aligned with critiques that current autoregressive systems lack persistent world models, grounded planning, and self-updating representations [45]. The point of the Artificial Einstein Test is to pressure-test that critique empirically, not to settle it by rhetoric.

14 Discussion

GPT-1900 is best read neither as a triumph nor a failure. It is a measuring instrument under construction. Its suggestive outputs show that time-locked models can generate historically plausible conceptual moves. Its failures show that plausible prose, high-temperature sampling, and curated prompts are insufficient evidence for scientific understanding.

This is precisely the value of the first deployment. It forces the benchmark to mature. The next version must be less like asking a model to answer a history-of-physics exam and more like asking it to participate in a controlled reconstruction of scientific discovery.

Zahavy’s position supplies a serious mechanism hypothesis for that reconstruction: action-controllable physical simulation may help a system generate premises that text-only optimization does not. The Artificial Einstein Test supplies the missing adjudication. It can compare matched text-only and interactive conditions, hide the relevant consequence, expose the final candidate to an independent verifier, and count every failed trajectory. This design can support a narrower conclusion—that intervention causally improves one or more stages of the jump—without turning a null result into a permanent theorem about all language models.

The most important question is not whether an AI can say what Einstein said. It is whether it can do what Einstein did: select the right anomaly, abandon the wrong ontology, import the right mathematics, preserve the right empirical constraints, and make a theory vulnerable to nature. The jump is scientifically meaningful only if the landing survives contact with evidence the system did not already see.

15 Conclusion

The Artificial Einstein Test remains a useful and unusually stringent benchmark for scientific AI. After GPT-1900, however, the paper should no longer be framed merely as a proposal. It should be framed as the first systematic interpretation of a new benchmark class.

The lesson is clear. Time-locked LLMs make scientific rediscovery tests possible, and the abductive-jump thesis identifies the stage current systems may be missing. But a passing test must require more than historical role-play, equation emission, conceptual fluency, or an impressive simulation. It must require temporally isolated evidence, robust evaluation, complete failure accounting, stage-gated prior selection, withheld prediction, and an independent discriminating test.

If future systems can pass that standard on general relativity, the result would be profound. If they cannot, the failure will be equally informative. Either outcome would move AGI evaluation away from imitation and toward the only standard that ultimately matters in physics: contact with reality.

References

- [1] A. Einstein, “Die Feldgleichungen der Gravitation,” *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin*, 844–847 (1915).
- [2] A. Einstein, *Autobiographical Notes*, in P. A. Schilpp, ed., *Albert Einstein: Philosopher-Scientist*, Open Court (1949).
- [3] B. Keating and L. Krauss, “Lawrence Krauss: The Physics of Everything,” *Into the Impossible*, July 30, 2021, discussion beginning at 2:18:02, <https://www.youtube.com/watch?v=mumFmyJtfNo&t=8282s>.
- [4] B. Keating and L. Krauss, “Artificial Intelligence: AI Physicists are INEVITABLE!,” August 5, 2021, discussion beginning at 2:23, <https://www.youtube.com/watch?v=6PBQg8I3a9Q&t=143s>.
- [5] B. Keating and E. Topol, “AI Doctors, Medicine’s Future, and the Delta Variant,” *Into the Impossible*, August 10, 2021, discussion beginning at 17:04, https://www.youtube.com/watch?v=DYqGmFY_BNI&t=1024s.
- [6] B. Keating and D. Chalmers, “The MATRIX was a DOCUMENTARY!,” *Into the Impossible*, February 9, 2022, <https://www.youtube.com/watch?v=ywjbbQXAFic>.
- [7] B. Keating, R. Penrose, and S. Hameroff, “Collapsing a Theory of Quantum Consciousness? Part 2,” *Into the Impossible*, August 5, 2022, <https://www.youtube.com/watch?v=0oDi856wLPM>.
- [8] B. Keating and N. Bostrom, “Superintelligence and the Simulation Hypothesis,” *Into the Impossible*, September 7, 2022, <https://www.youtube.com/watch?v=X0SssEwsu0A>.
- [9] B. Keating, “Albert Einstein’s HAPPIEST Thought,” December 31, 2023, https://www.youtube.com/watch?v=0bQ8-5ENo_w.
- [10] B. Keating and S. Wolfram, “My Discovery Changes Everything,” *Into the Impossible*, January 19, 2024, <https://www.youtube.com/watch?v=c-R03vM100k>.

- [11] B. Keating and M. Tegmark, “Will AI Surpass Human Intelligence?,” *Into the Impossible*, December 10, 2024, benchmark comparison beginning near 14:09, <https://www.youtube.com/watch?v=YwC16DhtkI&t=849s>.
- [12] B. Keating and Y. LeCun, “AI Doomsday Fears Are Overblown,” *Into the Impossible*, December 29, 2024, <https://www.youtube.com/watch?v=u7e0YUcZYbE>.
- [13] B. Keating and T. Sejnowski, “How to Train ChatGPT to Serve You,” *Into the Impossible*, January 19, 2025, discussion beginning near 12:20, https://www.youtube.com/watch?v=qP_ZeNyorkM&t=740s.
- [14] B. Keating, “This AI Broke Every Benchmark—Then It Did Something Worse,” *Into the Impossible*, April 16, 2026, transcript mirror, <https://podscripts.co/podcasts/into-the-impossible-with-brian-keating/this-ai-broke-every-benchmark-then-it-did-something-worse>.
- [15] B. Keating, “Free Fall to Faith: Einstein’s Elevator and Spinoza’s God,” *My Elevator Pitch for God*, <https://www.myelevatorpitchforgod.com/all-authors/brian-keating-ph.d.->, accessed August 18, 2026.
- [16] D. Hassabis, public remarks on a possible AGI test involving a model trained with a 1911 knowledge cutoff and asked to derive general relativity, reported in 2026 interviews and technology press coverage.
- [17] M. Hla, *GPT-1900*, GitHub repository, <https://github.com/michaelhla/gpt1900>, accessed June 7, 2026.
- [18] M. Hla, “Machina Mirabilis,” author essay, March 2026, <https://michaelhla.com/blog/machina-mirabilis.html>, accessed June 7, 2026.
- [19] Entanglr, *Awesome Vintage LLMs*, GitHub repository, <https://github.com/entanglr/awesome-vintage-llms>, accessed June 7, 2026.
- [20] D. Benrimoh, N. Mikus, P. Stone, and A. Rosenfeld, “The Einstein Test,” *Communications of the ACM* (2026), <https://dl.acm.org/doi/10.1145/3771100>.
- [21] L. Rosenberg, “Have we seen the last Einstein?,” *Big Think*, July 20, 2026, <https://bigthink.com/science-tech/have-we-seen-the-last-einstein/>.
- [22] Typewriter AI, *Typewriter-1913-7B-base*, Hugging Face model card, <https://huggingface.co/typewriter-ai/typewriter-1913-7B-base>, accessed July 18, 2026.
- [23] F. Chollet, M. Knoop, G. Kamradt, B. Landers, and H. Pinkard, “ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems,” arXiv:2505.11831 (2025).
- [24] J. Land, “Modality-Driven Search with Holistic Trace Judging for ARC-AGI-2,” arXiv:2606.31543 (2026).
- [25] A. J. Lew, Y. Cao, and M. J. Buehler, “ProjectionBench: Evaluating Scientific Hypothesis Generation in LLMs Under Progressive Information Disclosure,” arXiv:2605.30284 (2026).
- [26] C. Shi et al., “Are LLMs Ready for Scientific Discovery? A Capability-Oriented Benchmark for AI Scientists,” arXiv:2607.11079 (2026).

- [27] T. Miao et al., “PRL-Bench: A Comprehensive Benchmark Evaluating LLMs’ Capabilities in Frontier Physics Research,” arXiv:2604.15411 (2026).
- [28] M. Nagele and F. Marquardt, “Agentic Exploration of Physics Models,” *Physical Review X* **16**, 031002 (2026).
- [29] M. L. Wiemann, L. M. Smith, P. Melchior, S. Mishra-Sharma, A. G. Wilson, P. Izmailov, and C. Cuesta-Lazaro, “DiscoverPhysics: Benchmarking LLMs for Out-of-the-Box Scientific Thinking,” arXiv:2605.26087 (2026).
- [30] D. Zhang, “Testing Frontier Large Language Models’ Physics Literacy in Parallel Physical Worlds,” arXiv:2607.00276 (2026).
- [31] W. Xu et al., “ResearchClawBench: A Benchmark for End-to-End Autonomous Scientific Research,” arXiv:2606.07591 (2026).
- [32] T. Zahavy, “Position: LLMs Can’t Jump,” ICML 2026 Position Paper Track (2026), <https://www.tomzahavy.com/files/llms-cant-jump.pdf>, accessed August 18, 2026.
- [33] L. Magnani, *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*, Springer (2009).
- [34] J. Bruce et al., “Genie: Generative Interactive Environments,” arXiv:2402.15391 (2024).
- [35] J. D. Norton, “Einstein’s Pathway to General Relativity,” University of Pittsburgh (2020), https://sites.pitt.edu/~jdnorton/teaching/HPS_0410/chapters/general_relativity/index.html.
- [36] A. Karpathy, *nanochat*, GitHub repository, <https://github.com/karpathy/nanochat>.
- [37] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” *Nature* **529**, 484–489 (2016).
- [38] J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” *Nature* **596**, 583–589 (2021).
- [39] M. Schmidt and H. Lipson, “Distilling free-form natural laws from experimental data,” *Science* **324**, 81–85 (2009).
- [40] M. M. Bronstein, J. Bruna, T. Cohen, and P. Velickovic, “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges,” arXiv:2104.13478 (2021).
- [41] V. G. Satorras, E. Hoogeboom, and M. Welling, “E(n) Equivariant Graph Neural Networks,” *Proceedings of ICML* (2021).
- [42] F. Chollet, “On the Measure of Intelligence,” arXiv:1911.01547 (2019).
- [43] T. S. Kuhn, *The Structure of Scientific Revolutions*, University of Chicago Press (1962).
- [44] K. Popper, *The Logic of Scientific Discovery*, Routledge (1959).
- [45] Y. LeCun, public lectures and essays on objective-driven AI, world models, and the limitations of autoregressive language models.