Preprint. September 2, 2026.

Rediscovering Relativity Is Not a Single Task

What It Takes for AI to Discover a Physical Theory

Dong Zhang

Can AI rediscover general relativity? This question is increasingly discussed in the AI community, but it is often framed too narrowly. Some framings treat rediscovery as reproducing Einstein's field equations, others assume that a theory can be built from a single batch of experimental data, and still others reduce scientific discovery to one mode of inference. We argue that rediscovering a physical theory is not a single task, but a layered sequence of distinct capabilities. Using the history of relativity as a case study, we reconstruct it not as a straight path from Newton to Einstein, but as a branching theory space with alternatives from Lorentz, Poincaré, Nordström, and Einstein. We show that theory construction and experiment design were mutually generative rather than sequential, and that at several points the available evidence could not decide between competing theories. We therefore replace the idea of asking an AI system to jump directly to Einstein's final theory with three increasingly demanding tests. These tests measure physics intelligence level by level. They ask whether an AI system can revise a theory inside an old framework, replace patches with a principle, and construct candidate theories whose consequences can be tested.

1. Introduction

Can AI rediscover general relativity? This question has recently become a benchmark for evaluating the scientific capabilities of AI systems (Wang et al., 2023). One influential proposal is Hassabis's "Einstein test": train a system whose knowledge is restricted to what was known before 1911 and ask whether it can independently rediscover general relativity, just as Einstein did in 1915. Only then, the argument goes, could we say that the system has genuine general intelligence.[1] One such attempt trains models from scratch on corpora containing only pre-1900 texts, aggressively removing any traces of later concepts (Hla, 2026).

The motivation behind these efforts is sound: they try to hide the answer and reduce data contamination. But they still leave the central question underspecified. What counts as "rediscovering" relativity: Einstein's field equations, the route that leads to them, or one decisive step along the way? How should success or failure be judged before the end is reached? And does the answer even have to be Einstein's? More than one theory was consistent with everything known in 1911, and an AI system could reason faultlessly and arrive at one that later measurement ruled out. This paper addresses these questions. In our view, rediscovering relativity is not a single task; it is the reconstruction of a historical theory space. It follows that the physics intelligence being tested comes in levels, not as a single yes-or-no property.

More importantly, the current discussion often makes the history of discovering relativity look too static, as if all relevant facts had already been collected and the task were to infer the right theory from a fixed archive. The actual history was not shaped in this way. Maxwell's equations predicted electromagnetic waves and identified light as one of them. In the Newtonian framework, such waves were understood to propagate through the aether, which served as an absolute rest frame. This in turn motivated the search for the so-called "aether wind". The null result of the Michelson–Morley experiment (Michelson & Morley, 1887) led Lorentz to propose the contraction hypothesis, still within the Newtonian framework (Lorentz, 1895). The experiments carried out between 1902 and 1904 were no longer testing for the aether wind, but the new predictions generated by the contraction hypothesis itself (Rayleigh, 1902; Trouton & Noble, 1903; Brace, 1904). The history of relativity was an interaction between theory and experiment: theories suggested what to test next, and experiments forced theories to change. From Maxwell to Lorentz, from Poincaré to Einstein, and later to the competing theories of relativistic gravity, the development of relativity was a continuous process of mutual interaction rather than a one-shot inference.

Instead of discussing rediscovery as one big question, we turn it into three concrete discovery questions, each closer to the physics and to the historical details:

- Can AI construct and revise a theory inside an existing

dongzhanghz@gmail.com

[1] Demis Hassabis, remarks in "Fireside Chat: Sir Demis Hassabis and Prof Govindan Rangarajan," Indian Institute of Science, 17 February 2026, moderated by Varun Mayya. Transcript available at the IISc EECS event page.

framework as new experimental evidence arrives, as Lorentz did?
- Can AI stop patching the old framework and replace accumulated fixes with a unifying principle, as Poincaré and Einstein did?
- Can AI construct the entire space of candidate relativistic theories of gravity, derive their predictions, and identify the experiments that could distinguish among them?

These questions are steps rather than three versions of the same capability. The first works inside a framework it is given, keeping one theory in contact with experiments. The second gives up the repairs and replaces them with a principle that reaches beyond the scope of the existing framework. The third goes beyond that, developing new physics with falsifiable predictions. Together they define three ascending levels of AI physics intelligence rather than a single pass-or-fail test.

Our position is that rediscovering relativity involves several tasks rather than one, and that each of them sits at a different level of physics intelligence. It is not a matter of induction, deduction, or a single act of abduction, since the levels call on different modes of reasoning. Nor can it be completed on paper: new physics has to be proposed and then validated by experiments, and the final answer comes from the world rather than from the theory itself.

Section 2 reconstructs the theory space from the history, and Section 3 reviews three other framings. Section 4 decomposes rediscovery into distinct capabilities, and Section 5 turns them into a three-level scale with the tests that define each level.

2. The Historical Theory Space of Relativity

2.1. 19th-Century Situations

Newtonian mechanics and universal gravitation had won enormous authority in 19th-century astronomy. One good example was Neptune: irregularities in the motion of Uranus were used to infer the position of an unseen planet, which was then observed. Mercury's anomalous perihelion precession (Le Verrier, 1859) was a problem, but not yet a reason to abandon Newtonian gravity. The residual was about $43''$ per century, out of a total precession of roughly $5557''$.

The first responses were patches inside the Newtonian picture. Le Verrier proposed an intra-Mercurial planet, Vulcan, in 1859 (Le Verrier, 1859), and astronomers searched for it for decades without confirmation. Other fixes included solar oblateness and Hall's 1894 modification of the inverse-square exponent (Hall, 1894). These proposals had weaknesses, but Mercury alone did not force Einstein's answer or open up new physics. As long as old models still had plausible adjustments, a patched Newtonian theory remained the natural first move.

Electrodynamics created a different situation. Maxwell's equations unified electricity, magnetism, and light by treating light as an electromagnetic wave (Maxwell, 1865). But the equations were not covariant under Galilean transformations, which made the aether look like the natural rest frame for electromagnetic waves (Maxwell, 1880). The Michelson–Morley experiment (Michelson & Morley, 1887) was designed to detect the aether wind, the motion of the Earth through this frame. Its result was null. By the end of the 19th century, the connection between Newtonian mechanics and Maxwell's electrodynamics was under strain. The aether frame was what allowed the two to coexist, and Michelson–Morley found no sign of it. Relativity grew out of this tension.

2.2. Special Relativity: Lorentz, Poincaré, Einstein

Lorentz's theory lived inside the aether framework. His 1895 theory (Lorentz, 1895) used local time and length contraction to explain why the Michelson–Morley experiment did not detect aether drift. Later Rayleigh (Rayleigh, 1902), Brace (Brace, 1904), and Morley–Miller (Morley & Miller, 1905) tested consequences of the FitzGerald–Lorentz contraction idea (FitzGerald, 1889); Trouton–Noble tested whether a charged capacitor moving through the aether should feel a torque (Trouton & Noble, 1903). Lorentz had to update his theory with a more elaborate electron model and additional stabilising assumptions to match these new experiments (Lorentz, 1904), but he did not abandon the aether. In this sense, Lorentz was still patching a physical picture.

Compared with Lorentz's increasingly complex theory, Poincaré started from a simpler first principle (he called it "the principle of relativity"). In 1904, he treated the impossibility of detecting aether drift as a general postulate (Poincaré, 1904):

> *The laws of physical phenomena should be the same, whether for an observer fixed, or for an observer carried along in a uniform movement of translation; so that we have not and could not have any means of discerning whether or not we are carried along in such a motion.*

He then corrected Lorentz's formulation, showed that the Lorentz transformations form a group (Poincaré, 1906), and made the symmetry requirement explicit. Physically, Poincaré still accepted the aether framework and treated the time and length measured in other frames as artificial quantities. Mathematically, this brought his theory very close to Einstein's.

Unlike the other two, Einstein removed the privilege of

the aether and the absolute frame. His 1905 work (Einstein, 1905a) started from the relativity principle and the constancy of the speed of light, without referring to any hidden rest frame. He stated the relativity principle in terms very close to Poincaré's:

> *The same laws of electrodynamics and optics will be valid for all frames of reference for which the equations of mechanics hold good.*

Poincaré and Einstein used the same mathematics but gave it different physical meanings. Poincaré kept the aether and a true state of rest, so absolute velocity remained in the theory, even though no measurement could detect it. In Einstein's theory, only measurable relative velocities exist, so absolute velocity disappears. But in 1905 no experiment could decide between them. For the experiments then at issue, Lorentz, Poincaré, and Einstein gave the same observable predictions. The mass–energy question does not change this point. See Section B.

2.3. Theory and Experiment as Mutually Generative

We clarify the interaction between theory and experiment. Between 1887 and 1902 there was only one aether-drift experiment, and Lorentz's 1895 theory was good enough to account for it. The experiments that followed were of a different kind: they tested predictions that came from the contraction hypothesis, not the aether wind. All of them were null, and they pushed Lorentz from contraction alone to the full transformation and a more elaborate electron theory. Theory and experiment were driving each other inside the classical framework. What came next was different: Poincaré and Einstein left that framework, and no observation selected either of them.[2]

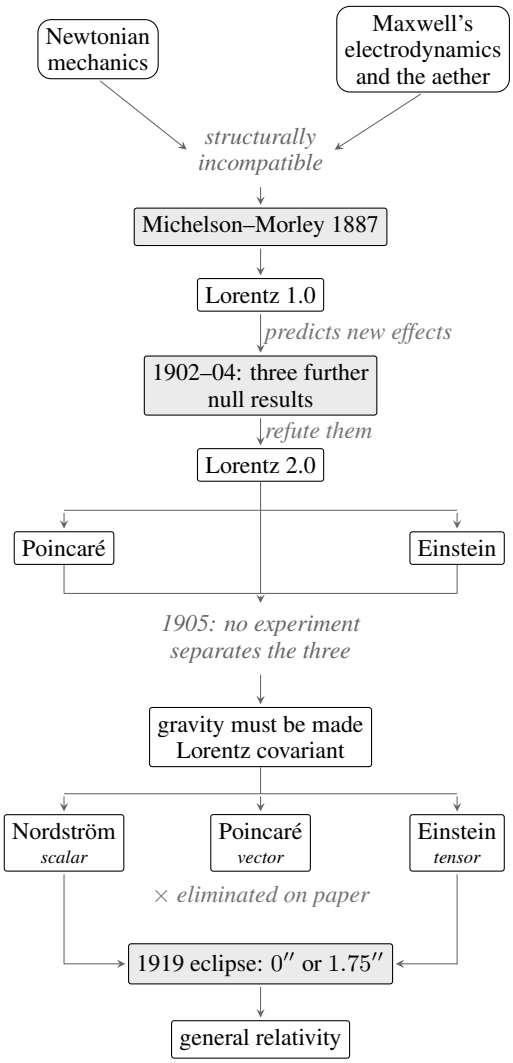


Figure 1. The theory space from classical physics to general relativity, with shaded boxes as the key experiments.

2.4. Relativistic Gravity Before General Relativity

Special relativity is naturally described in Minkowski spacetime (Minkowski, 1908), but Newtonian gravity does not fit there. Poisson's equation is not Lorentz covariant. How should gravity be brought into Minkowski spacetime? Between 1905 and 1915 this was an open problem worked on by Poincaré, Nordström, Abraham and Mie, not by Einstein alone (Walter, 2007).[3] The standard move was to keep spacetime flat and add a gravitational field on it, exactly as one adds an electromagnetic field. The field has to be Lorentz covariant, it has to reduce to Newtonian gravity in the static weak-field limit, and its source has to be the energy-momentum tensor $T_{\mu\nu}$.

Poincaré took the vector route in 1905, modelling gravity on electromagnetism (Poincaré, 1906; Walter, 2007), where the source is a four-current j^μ and the field a four-potential A^μ. It was the only template available, since the energy-momentum tensor did not yet exist as a tool (von Laue, 1911). It made gravity propagate at speed c and included what Poincaré called gravitational waves, but little else to

[2] Quantum theory shows the pattern even more plainly. The move from classical mechanics to quantum mechanics, and later to quantum field theory, did not come from a few isolated observations. It came from a long back-and-forth between theory and experiment, where each side repeatedly changed what the other could ask. The path to quantum field theory was built through that loop, not derived on paper.

[3] Abraham and Mie are not ignored because they were historically irrelevant. If an AI system produced Abraham's theory, that would count as generating a real historical branch, but it should also be able to reject it on consistency grounds (Einstein, 1912). If it produced Mie's programme, that would show a historically plausible search direction, but one outside the scalar–vector–tensor comparison below unless it yielded clear gravitational predictions (Smeenk & Martin, 2007).

test with observation: a perihelion advance for Mercury well short of the observed $43''$, and no other prediction.

Nordström kept Newton's potential as a scalar and kept it as a field on Minkowski spacetime (Nordström, 1913; Norton, 1992). Newton's equation contains derivatives in space only, which is why its gravity acts everywhere at once. Nordström added the missing time derivative, which turns it into a wave equation and makes the potential travel at c. In its mature form, the field equation was

$$\phi\Box\phi = -4\pi GT,$$

where ϕ is the scalar gravitational field, the relativistic version of Newton's gravitational potential, and $T = T^{\mu}{}_{\mu}$ is the trace of the energy-momentum tensor. The source is a scalar because the gravitational field is scalar. His theory was self-consistent and attractive, and it explicitly requires $m_i = m_g$, so the equivalence principle was not Einstein's private concern, and the gravitational redshift comes out right for that reason. However, the theory predicts a perihelion precession of $-7''$ per century for Mercury, with the sign reversed, and no bending of light passing through the gravitational field of matter.

Einstein also began with a scalar, but for a different reason. From 1907 to 1912, the equivalence principle led him to describe gravity through a position-dependent speed of light $c(x)$, which gave the $0.83''$ light deflection he published in 1911 (Einstein, 1911). This was close to Nordström's scalar route in form, but not in motivation: Einstein was trying to make gravity express the local equivalence of acceleration and gravitation, and the scalar could not carry that idea beyond static fields. The turn came in 1912–1913, when Grossmann supplied the tensor calculus and Riemannian geometry needed to treat gravity as the metric of spacetime itself. The Entwurf, the outline theory he wrote with Grossmann (Einstein & Grossmann, 1913), was already a metric theory, but its field equations were still wrong: they were not fully generally covariant, and they gave only $18''$ for Mercury. Einstein rebuilt the equations in November 1915 (Einstein, 1915a;b),

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu},$$

where $R_{\mu\nu}$ is the Ricci curvature tensor, R is the scalar curvature, and $g_{\mu\nu}$ is the spacetime metric. Only then did the predictions come out right: $+43''$ per century for Mercury and $1.75''$ for light grazing the Sun.

The lesson is not that Einstein's path was the only path to the theory. Poincaré, Nordström and Einstein moved in different ways. What they shared was the same post-1905 problem: how to make gravity compatible with relativity. That problem opened scalar, vector and tensor answers. Observation separated them and selected Einstein's answer.

3. Other Framings of Rediscovery

3.1. The Whig History Framing

Whig history (Butterfield, 1931) writes the past from the "winning end". It keeps the line that leads to the present, drops the branches that did not survive, and makes the outcome look inevitable.

Applied to relativity, it records Einstein and his reasoning alone, and omits the parallel work of Lorentz, Poincaré, Nordström, Abraham, Mie and so on. In this version, relativity becomes a straight Einstein-centered story: special relativity rests on two postulates, the principle of relativity and the constancy of the speed of light, while general relativity begins from Mercury's anomalous perihelion, passes through the equivalence principle and Riemannian geometry, and arrives at Einstein's field equations.

As a framing of rediscovery, it carries a further assumption: at each historical moment, there was already one definite correct answer, and rediscovery means reaching it again. This assumption makes the framing easy to turn into a test. An AI evaluation built on it can score an AI system by asking whether it outputs the two postulates of special relativity, or the Einstein field equations, and count anything else as failure. This framing appears in the Einstein test quoted in Section 1, in benchmark designs that train on pre-1900 corpora and then check for traces of later concepts (Hla, 2026), and in popular accounts that follow Einstein alone.

3.2. The One-Shot Framing

On this view, once the results of the key experiments are in place, the theory follows from them. Discovery is therefore treated as a processing step: the work starts from a fixed archive and turns it into a theoretical result (Brunton et al., 2016; Udrescu & Tegmark, 2020).

Applied to relativity, this account says the following. By the turn of the century, the aether-drift experiments were already in hand, and special relativity follows from them. Mercury's anomalous perihelion precession was also already known from 19th-century astronomy. Once special relativity is available, that old discrepancy becomes the remaining input from which general relativity follows. Each theory appears in a single step, as the correct reading of the evidence available at the time.

Moved into AI, the framing gives a clean experimental design: give an AI system all of physics up to 1903 and have it produce special and general relativity. It appears in public statements by the leadership of AI laboratories, and in benchmark designs that supply a complete corpus in a single prompt and then check whether the target theory is present in the output (Hla, 2026; Lu et al., 2024). It presupposes that the evidence base is fixed before theorising begins, so

that all of discovery can be placed in one inference from data to theory.

3.3. The Single-Mechanism Framing

On this view, a theory emerges once the right method is applied. One operation turns what is known into a theory, and different accounts name different methods: induction, deduction, abduction, compression, or search. Each treats the operation it names as the whole task.

Applied to relativity, the episode becomes a case where a single mode of inference worked. In the inductive reading, Einstein found the shortest description covering the observations, and the theory is the compressed form of the data. In the deductive reading, Einstein found the right postulates and derived the theory from them. In the abductive reading, Einstein made one leap to the best explanation of the available evidence. The later work only developed that explanation further.

Each reading also has a modern AI version. Induction often appears as compression: discovery means finding the shortest description of the data (Schmidhuber, 2009). Deduction appears in formal mathematics and systems such as AlphaProof (Hubert et al., 2025), where results are found by searching a fixed set of axioms. Abduction appears as the claim that discovery needs one creative jump to a new explanation (Zahavy, 2026). The shared idea is simple: improve that one mechanism, and discovery improves with it.

A more detailed chronological version of these changes is given in Section A.

4. Rediscovery Is Not One Task

Two facts often placed at the origin of relativity did not by themselves produce it. The equality of inertial and gravitational mass ($m_i = m_g$) was a constraint that any viable theory had to respect, not yet the problem that drove the theory. Mercury's unexplained $43''$ per century was real, but Newtonian fixes could still absorb it. The stronger pressure came from electrodynamics: Maxwell's theory did not fit Galilean relativity, and the expected aether wind was not found.

That incompatibility opened the problem rather than settling it. What followed took sixty years, and it was not a single step. It included an explanation of the null result still inside the old framework, revision under a second round of experiments, a new principle, special relativity, the separate problem of making gravity Lorentz covariant, a space of candidate theories, and finally an observation that chose among them. Table 1 sets out the task chain.

Table 1 has one central lesson. Relativity was not a task that could be completed by theory alone. The shaded rows are experiments, and their answers had to come from the world. The rest are theory, but each depends on earlier experimental results, and several of them leave more than one candidate alive. The task is therefore a chain: theory proposes, experiment tests, theory revises, and only then does the next theoretical move become possible.

The implication for AI is direct. If an AI system is asked to rediscover relativity, it should not be judged as if the task were to produce the final theory in one step. It can do the theoretical work: recognise problems, build candidate theories, derive predictions, and design the experiments that would separate them. But the full loop cannot be closed inside the system. At several points, parallel theories fit the evidence already available, and only experiment could decide among them. AI rediscovery should therefore be staged and world-coupled, not an automatic derivation from a fixed archive.

5. Three Levels of Physics Intelligence

Section 4 broke the rediscovery problem into a chain of tasks. We can now turn that chain into evaluations. We describe three tests, each built from one moment in the chain: Lorentz's revision under new experiments, Poincaré's move from Lorentz's patches to a principle, and the construction of the space of relativistic theories of gravity. They are not three independent probes but three levels of physics intelligence, labelled Level 1 to Level 3. Graded scales have been used for general AI capability (Morris et al., 2024), here we use the same idea for the construction of physical theory. Each test has the same structure: we say what is supplied, what is withheld, what the system is asked to do, and what would count as success.

These are conceptual designs rather than benchmarks ready to run. Every answer is present in any modern training corpus, so an AI system that simply outputs Einstein's postulates has not shown discovery rather than recall.

5.1. The Lorentz Test

The first test gives the evidence in stages, because the order matters. The AI system receives Maxwell's equations, the aether framework, and the null result of Michelson–Morley. The experiments of 1902 to 1904 are withheld, along with the Lorentz transformation, the relativity principle, and everything after them.

It is asked to explain the null result, and then to derive what else that explanation implies. If lengths contract along the direction of motion, what else should be visible, and in what apparatus? Only once those predictions are recorded does the second batch arrive (as discussed above in Section 2): Rayleigh and Brace tested double refraction in moving matter; Morley and Miller tested whether the effect depends

#	Task	Capability it requires	Can AI attempt it?
0	Newtonian mechanics and Maxwell's electrodynamics	starting point	—
1	Michelson–Morley: measure the speed of the aether wind	turning a theoretical question into an apparatus	design only
2	Lorentz 1.0: contraction and local time	theory construction on incomplete evidence	yes (the Lorentz test I)
3	Rayleigh, Brace, Morley–Miller, Trouton–Noble	deriving new testable questions from a theory	design only
4	Lorentz 2.0: the full transformation	revision under a second round of null results	yes (the Lorentz test II)
5	Poincaré: the relativity principle	principle formation from accumulated patches	yes (the Poincaré test)
6	Einstein: special relativity	reformulation without a privileged frame	yes (the Poincaré test)
7	Pose the next problem: make gravity Lorentz covariant	problem recognition	yes (the gravity test)
8a	Scalar, vector and tensor candidates	hypothesis-space construction	yes (the gravity test)
8b	Predictions for each candidate	derivation and screening on paper	yes (the gravity test)
9	Deflection of starlight measured at a solar eclipse, 1919	identifying the discriminating observation	design only
10	Gravitational redshift confirmed, 1959	later confirmation, does not discriminate	design only
11	Gravitational waves detected, 2015	later confirmation	design only

Table 1. The relativity episode decomposed into tasks. Shaded rows are the tasks that have to be settled in the world. "Design only" means an AI can specify experiments, down to apparatus and required precision, but humans have to build them and run them, and the answer comes from the world rather than from the reasoning.

Rank	Field	Couples to	Mercury precession	Light deflection	Redshift	Outcome
0	scalar φ	T^{μ}_{μ}, the trace	$-7''$, sign reversed	exactly zero	correct	requires observation
1	vector A_{μ}	mass current	not computed	not computed	not computed	eliminated on paper
2	tensor $h_{\mu\nu}$	$T_{\mu\nu}$	$+43''$	$1.75''$	correct	requires observation

Table 2. Three candidates in the gravity test (Section 5.3), ordered by the rank of the field. The vector branch is eliminated on theoretical grounds before the observables are computed; the shaded rows are the candidates that survive on paper and require observation.

on the material the interferometer is built from; Trouton and Noble tested the torque expected on a moving charged capacitor. All three results are null. The system must then decide whether to patch its hypothesis, abandon it, or rebuild.

Success here is not simply writing down the Lorentz transformation. It is explaining the new results inside the old framework. The system has to derive the predictions that the second batch will test from its own hypothesis, recognise that those predictions have been refuted, and say what the repair has cost in added assumptions. Staying inside the old framework is enough here. The next level requires moving beyond it. This is the capability listed against tasks 2 and 4 in Table 1. It cannot be assessed if the two batches arrive together.

5.2. The Poincaré Test

The second test supplies no new evidence at all. The AI system receives Maxwell's equations and the whole set of null results at once, and is asked to reorganise what it already has. Withheld are the relativity principle, the Lorentz group and Einstein's postulates. This is where economy of thought enters the test. The point is not to add one more device that fits another null result, but to find one rule that makes the separate devices unnecessary. The difficulty here is therefore not revision but elevation: a set of local patches has to give way to a unifying principle.

Failure means giving a better patch in the old framework, not a principle in a new framework. An AI system may return a more elaborate electron model that accounts for every null result, but each result is still handled separately. Success means completing the whole move: turning the null results into one general rule, turning that rule into a relativity principle, deriving the Lorentz transformations, and showing that the transformations form a group. The group result matters because any change from one inertial frame to another stays inside the same rule. The first rule may still be in aether language: no experiment can detect uniform motion through the aether. The principle may keep the aether, as in Poincaré, or remove it, as in Einstein. We judge whether the AI system has replaced separate fixes

Level	Test	Counts as success	Counts as failure
1	The Lorentz test	explains the new results while keeping the old framework, by adding another fix instead of giving it up	produces a hypothesis with no consequences of its own, so the second batch has nothing to refute
2	The Poincaré test	leaves the old framework and states a principle, so that the accumulated patches collapse into one semi-axiomatic system	goes on patching inside the old framework: every result is fitted, and nothing further follows
3	The gravity test	recovers at least the scalar, vector and second-rank tensor branches, eliminates the vector on theoretical grounds, and identifies the benchmark observable that separates the survivors	offers only one candidate, or leaves the vector branch alive as an observation-requiring theory

Table 3. Three levels of AI physics intelligence proposed in Section 5. Each row states what counts as success and what kind of failure the test is designed to expose.

with one rule and one transformation law. What is scored is the organizing principle added to the framework, not the number of axioms in the final theory.

Nothing observable separated Poincaré's formulation from Einstein's in 1905, and scoring Einstein higher would turn the Whig framing of Section 3 into a grading rule. The advantage of Einstein's route appeared only later, at the gravity stage. Keeping the aether is therefore not an error in this test, but a branch whose cost had not yet come due. Extra credit goes to an AI system that notices the two formulations are observationally indistinguishable on the evidence available. This is the capability listed against task 5 in Table 1.

5.3. The Gravity Test

The third test does not simulate Einstein's historical route. It asks whether an AI system can reconstruct the candidate space of relativistic gravity, rather than merely reach the theory that eventually won. The starting point is the flat Minkowski spacetime of special relativity. Gravity is then introduced as an additional field on that background. Matter carries a symmetric second-rank energy-momentum tensor $T_{\mu\nu}$ obeying $\partial^{\mu}T_{\mu\nu} = 0$. The gravitational field is coupled to matter, and the theory reduces to Newtonian gravity in the static weak-field limit. The system is asked to enumerate the possible field types and say on what grounds, then to write the coupling and the field equations for each. Rather than being told which relativistic effects to calculate, it must derive the observable consequences needed to compare the candidates. It must then decide which candidates can be eliminated on paper, which can be separated only by experiment, and which experiment would separate them. The equivalence principle, Riemannian geometry, Einstein and the 1919 measurement are not mentioned.

Two kinds of answer fail. Producing a single candidate fails, whether it is the scalar that continues Poisson's equation or the tensor that happens to be right, because the space was never generated. Leaving the vector theory alive also fails, because a vector gravitational field would make like masses repel rather than attract.

Success is the content of Table 2. The scalar, vector and second-rank tensor branches are the minimum target, not a closed taxonomy. A stronger answer may propose additional candidates, but it must still recover these three historically and physically central cases. The system should make its own predictions for observable effects. For scoring, we use three benchmark observables: Mercury's perihelion precession, gravitational redshift, and solar light deflection. Gravitational redshift does not separate the scalar and tensor branches, since both give the effect. Mercury is a weak discriminator because auxiliary astronomical assumptions, such as solar oblateness or unseen mass, can move the target, and the anomaly was already known before the theories were built. Light deflection is decisive: the scalar branch predicts zero deflection, while the tensor branch predicts a nonzero effect that had not yet been measured.

What is scored is the route, not the final name of the theory. From this flat-background starting point, the tensor branch can lead through self-coupling to equations dynamically equivalent to Einstein's equations, as later field-theoretic derivations showed (Gupta, 1954; Kraichnan, 1955; Feynman et al., 1995; Deser, 1970). But simply invoking curved spacetime or naming general relativity does not satisfy the test, because those ideas are withheld.

The three tests now have to be read together. They are not a checklist of independent tasks, but three levels of physics intelligence, summarised in Table 3. Level 1 keeps a single theory alive under new evidence; Level 2 replaces accumulated patches with a principle; Level 3 constructs a candidate space and identifies the observation that would separate its survivors. The levels are not strictly nested, but their demands widen: from managing one theory, to reorganising a framework, to deciding what theory alone can settle and what must be asked of the world.

6. Conclusion

Can AI rediscover general relativity? The question does not take a simple yes-or-no answer. Historically, relativity was not a task that could be completed by theory alone. The episode was a chain: theory proposes, experiment tests, theory revises, and only then does the next theoretical move become possible. Several competing theories ran in parallel. Nothing inside the theories themselves could decide which one was right; the difference had to be settled by experiment (Figure 1 and Table 1). If AI is to rediscover relativity, the target is therefore not a single leap from Newton's physical picture to Einstein's. It requires a multi-step, multi-dimensional standard for measuring the physics intelligence of the AI system. This is why the usual framings are too narrow: the task is not to recover the winning theory, infer from a fixed archive, or reduce discovery to one mode of reasoning.

This paper breaks the question of whether AI can rediscover general relativity into three more concrete tests: the Lorentz test, the Poincaré test, and the gravity test (Table 3). The Lorentz test asks whether an AI system can work with experiments over time: explain a result inside an existing framework, derive new tests from its own hypothesis, and revise the theory when those tests come back. This is what Lorentz and others did. The Poincaré test asks for economy of thought in this precise sense: whether an AI system can turn a series of patches made inside the old framework into a unifying principle, even when that principle goes beyond the framework that produced the patches. This is what Poincaré and Einstein did, and it is how special relativity was built.

The third test, the gravity test, asks whether an AI system can make gravity compatible with special relativity. The target is not a direct jump to Einstein's version of general relativity, but a set of candidate theories: historically, Nordström's scalar theory, Poincaré's vector theory, and Einstein's tensor theory (Table 2). What is tested is whether an AI system can search the space of self-consistent constructions on its own, eliminate the candidates that fail on theoretical grounds, and propose the observations that could falsify what remains.

Note that in all three tests, the AI system is expected to know where its own limits are. An AI system working in mathematics can check its own output, and an AI system working in physics cannot; Section C sets out the contrast. At the special-relativity stage, there is no way to say whether Lorentz's, Poincaré's, or Einstein's physical picture is the right one, since all three account for the same phenomena to the same degree. The Lorentz test and the Poincaré test therefore probe the ability to construct theories, but do not ask the system to choose a winner. General relativity was not selected until the solar eclipse of 1919, so in the gravity test the system's task is to search the possible theories and propose the experiments, not to decide on theoretical or philosophical grounds which rival is superior.

The aim of this paper is to state a position and to lay out a sequence of increasingly demanding tests for probing the physics intelligence of an AI system. These tests are conceptual designs rather than benchmarks ready to run. We specify the inputs, the withheld information, and the success and failure criteria only at a high level (Table 3). The next step is to build evaluations in this form. They should not be prompts that ask for the final theory, but staged tasks that require an AI system to generate candidates, derive consequences, and name the experiment that would kill or separate what remains. We also do not claim that the three tests designed here can be transferred unchanged to other branches of physics, such as quantum mechanics, or to other sciences, such as chemistry and materials science.

References

Abbott, B. P. et al. Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*, 116 (6):061102, 2016.

Airy, G. B. On a supposed alteration in the amount of astronomical aberration of light, produced by the passage of the light through a considerable thickness of refracting medium. *Proceedings of the Royal Society of London*, 20: 35–39, 1871.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

Brace, D. B. On double refraction in matter moving through the aether. *Philosophical Magazine*, 7(40):317–329, 1904.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Butterfield, H. *The Whig Interpretation of History*. G. Bell and Sons, London, 1931.

Deser, S. Self-interaction and gauge invariance. *General Relativity and Gravitation*, 1(1):9–18, 1970.

Dyson, F. W., Eddington, A. S., and Davidson, C. A determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society A*, 220:291–333, 1920.

Einstein, A. Zur Elektrodynamik bewegter Körper. *Annalen der Physik*, 322(10):891–921, 1905a. English translation

by W. Perrett and G. B. Jeffery, *The Principle of Relativity*, Methuen, 1923.

Einstein, A. Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig? *Annalen der Physik*, 323(13): 639–641, 1905b. doi: 10.1002/andp.19053231314.

Einstein, A. Das Prinzip von der Erhaltung der Schwerpunktsbewegung und die Trägheit der Energie. *Annalen der Physik*, 325(8):627–633, 1906. doi: 10.1002/andp.19063250814.

Einstein, A. Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes. *Annalen der Physik*, 340(10): 898–908, 1911.

Einstein, A. Relativität und gravitation. erwiderung auf eine bemerkung von m. abraham. *Annalen der Physik*, 343 (10):1059–1064, 1912. doi: 10.1002/andp.19123431014.

Einstein, A. Die Feldgleichungen der Gravitation. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*, pp. 844–847, 1915a.

Einstein, A. Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*, pp. 831–839, 1915b.

Einstein, A. and Grossmann, M. *Entwurf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation*. B. G. Teubner, Leipzig, 1913.

Feynman, R. P., Morinigo, F. B., and Wagner, W. G. *Feynman Lectures on Gravitation*. Addison-Wesley, Reading, MA, 1995.

FitzGerald, G. F. The ether and the Earth's atmosphere. *Science*, 13(328):390, 1889.

Gupta, S. N. Gravitation and electromagnetism. *Physical Review*, 96(6):1683–1685, 1954.

Hall, A. A suggestion in the theory of Mercury. *The Astronomical Journal*, 14:49–51, 1894.

Hilbert, D. Die Grundlagen der Physik (Erste Mitteilung). *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*, pp. 395–407, 1915.

Hla, M. Machina mirabilis. `https://michaelhla.com/blog/machina-mirabilis.html`, March 2026. Online research note.

Hubert, T., Mehta, R., Sartran, L., Horváth, M. Z., Žužić, G., Wieser, E., et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651:607–613, 2025. doi: 10.1038/s41586-025-09833-y.

Hulse, R. A. and Taylor, J. H. Discovery of a pulsar in a binary system. *Astrophysical Journal*, 195:L51–L53, 1975.

Jumper, J., Evans, R., Pritzel, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596: 583–589, 2021.

Kraichnan, R. H. Special-relativistic derivation of generally covariant gravitation theory. *Physical Review*, 98(4):1118–1122, 1955.

Le Verrier, U. J. Lettre de m. Le Verrier à m. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 49:379–383, 1859.

Lorentz, H. A. *Versuch einer Theorie der electrischen und optischen Erscheinungen in bewegten Körpern*. E. J. Brill, Leiden, 1895.

Lorentz, H. A. Electromagnetic phenomena in a system moving with any velocity smaller than that of light. *Proceedings of the Royal Netherlands Academy of Arts and Sciences*, 6:809–831, 1904.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery, 2024.

Maxwell, J. C. A dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London*, 155:459–512, 1865. doi: 10.1098/rstl.1865.0008.

Maxwell, J. C. On a possible mode of detecting a motion of the solar system through the luminiferous ether. *Nature*, 21:314–315, 1880.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023.

Michelson, A. A. and Morley, E. W. On the relative motion of the Earth and the luminiferous ether. *American Journal of Science*, 34(203):333–345, 1887.

Mie, G. Grundlagen einer Theorie der Materie (Dritte Mitteilung). *Annalen der Physik*, 40:1–66, 1913.

Minkowski, H. Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pp. 53–111, 1908.

Morley, E. W. and Miller, D. C. Report of an experiment to detect the FitzGerald–Lorentz effect. *Proceedings of the American Academy of Arts and Sciences*, 41(12):321–328, 1905.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Nordström, G. Zur Theorie der Gravitation vom Standpunkt des Relativitätsprinzips. *Annalen der Physik*, 347(13): 533–554, 1913.

Nordström, G. On the energy of the gravitation field in Einstein's theory. *Koninklijke Nederlandsche Akademie van Wetenschappen Proceedings*, 20:1238, 1918.

Norton, J. D. Einstein, Nordström and the early demise of scalar, Lorentz covariant theories of gravitation. *Archive for History of Exact Sciences*, 45(1):17–94, 1992.

Poincaré, H. La théorie de Lorentz et le principe de réaction. *Archives néerlandaises des sciences exactes et naturelles*, 5:252–278, 1900.

Poincaré, H. L'état actuel et l'avenir de la physique mathématique. *Bulletin des Sciences Mathématiques*, 28:302–324, 1904. English translation by G. B. Halsted, *The Monist* 15, 1–24, 1905.

Poincaré, H. Sur la dynamique de l'électron. *Rendiconti del Circolo Matematico di Palermo*, 21:129–176, 1906.

Pound, R. V. and Rebka, G. A. Gravitational red-shift in nuclear resonance. *Physical Review Letters*, 3(9):439–441, 1959.

Rayleigh, L. Does motion through the aether cause double refraction? *Philosophical Magazine*, 4(24):678–683, 1902.

Schmidhuber, J. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems*, volume 5499 of *Lecture Notes in Computer Science*, pp. 48–76. Springer, 2009.

Schwarzschild, K. Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*, pp. 189–196, 1916.

Smeenk, C. and Martin, C. Mie's theories of matter and gravitation. In Renn, J. and Schemmel, M. (eds.), *The Genesis of General Relativity, Volume 4*, pp. 1543–1553. Springer, Dordrecht, 2007. doi: 10.1007/978-1-4020-4000-9_35.

Trouton, F. T. and Noble, H. R. The mechanical forces acting on a charged electric condenser moving through space. *Philosophical Transactions of the Royal Society A*, 202:165–181, 1903.

Udrescu, S.-M. and Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

von Laue, M. Zur Dynamik der Relativitätstheorie. *Annalen der Physik*, 340(8):524–542, 1911.

Walter, S. A. Breaking in the 4-vectors: The four-dimensional movement in Gravitation, 1905–1910. In Renn, J. and Schemmel, M. (eds.), *The Genesis of General Relativity, Volume 3*, pp. 193–252. Springer, Dordrecht, 2007.

Wang, H., Fu, T., Du, Y., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023.

Whitehead, A. N. *The Principle of Relativity with Applications to Physical Science*. Cambridge University Press, Cambridge, 1922.

Zahavy, T. Position: LLMs can't jump. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

A. Timeline of the Relativity Episode, 1859–2015

Figure 1 is the skeleton of what follows. Entries are given in the order they happened, with no separation between theory and experiment, because the point of the sequence is that the two alternate.

1859. Le Verrier reported an unexplained residual in Mercury's perihelion advance, about $38''$ per century in his calculation (Le Verrier, 1859). The value was later revised to about $43''$.

1859 onward. Vulcan, an intra-Mercurial planet, was proposed and searched for repeatedly. It was never confirmed.

1865. Maxwell's equations unified electricity, magnetism and light (Maxwell, 1865). The speed c entered as a constant of the medium, which made a preferred frame look natural.

1871. Airy repeated the aberration measurement with a water-filled telescope. No first-order effect of motion through the aether appeared (Airy, 1871).

1880. Maxwell's letter to D. P. Todd, published after his death, asked how the Earth's motion through the aether might be measured, and noted that any terrestrial round-trip method would need precision of second order in v/c (Maxwell, 1880). Michelson read it and set out to reach that precision.

1881. Michelson's first interferometer trial at Potsdam gave a negative result, but later corrections to the expected shift made the experiment inconclusive.

1887. Michelson and Morley performed the second-order experiment. A fringe shift of about 0.4 was expected; less than 0.01 was observed (Michelson & Morley, 1887).

1889. FitzGerald suggested that a body contracts along its direction of motion (FitzGerald, 1889).

1892–1895. Lorentz arrived at the contraction independently, arguing that if molecular forces were electromagnetic they should deform by the same factor, and added local time $t' = t - vx/c^2$, which he presented as a mathematical auxiliary (Lorentz, 1895).

1894. Hall proposed replacing the exponent in the inverse-square law by 2.00000016, which absorbed Mercury's residual without touching anything else (Hall, 1894).

1898. Poincaré argued that the simultaneity of distant events was a convention rather than a fact.

1900. Poincaré reinterpreted Lorentz's local time as what clocks synchronised by light signals would actually read (Poincaré, 1900).

1901–1906. Kaufmann measured the deflection of β rays. The results were read at the time as favouring Abraham's rigid electron over Lorentz's deformable one. This was the only place in the period where experiments appeared to be deciding between electron theories.

1902. Rayleigh looked for the double refraction that contraction should have produced in moving matter. Null (Rayleigh, 1902).

1903. Trouton and Noble looked for the torque that should have acted on a charged capacitor moving through the aether. Null (Trouton & Noble, 1903).

1904. Brace repeated the double-refraction search at higher sensitivity. Null (Brace, 1904).

1904. Lorentz published the full transformation, together with a deformable electron and further non-electromagnetic assumptions to hold it stable (Lorentz, 1904).

1904. At St Louis, Poincaré named the relativity principle and stated it as a postulate: no experiment can detect uniform motion (Poincaré, 1904). The arrow was now reversed. Lorentz had assumed the contraction and deduced that nothing could be measured; Poincaré assumed that nothing could be measured and deduced that the contraction had to occur.

1905. Morley and Miller varied the material the interferometer was built from. Null (Morley & Miller, 1905).

1905. Einstein derived the transformation from two postulates, discarded the aether, and treated t' as the time of the moving frame rather than a bookkeeping device (Einstein, 1905a). In later accounts, Einstein treated Fizeau's experiment and stellar aberration as especially important; neither was a null-result experiment.

1906. Poincaré's Palermo memoir proved that the transformations form a group. Its final section proposed a Lorentz-invariant law of gravitation propagating at c, and named gravitational waves (Poincaré, 1906). The extra perihelion advance it yielded fell well short of $43''$, which Poincaré reported as consistency with observation rather than as an explanation.

1907. Einstein stated the equivalence principle: an observer in free fall does not feel his own weight.

1908. Minkowski gave special relativity its four-dimensional form and wrote down the electromagnetic energy-momentum tensor (Minkowski, 1908). He also proposed a gravitation theory using four-dimensional retarded potentials.

1908. Bucherer's measurements reversed the verdict of the Kaufmann experiments. The rigid electron was abandoned.

1909. Ehrenfest posed the rotating disc: the circumference contracts and the radius does not, so the geometry of a rotating frame is not Euclidean.

1911. Einstein computed the deflection of starlight from the equivalence principle alone and obtained $0.83''$ (Einstein, 1911). Half of the effect was missing, because the contribution of spatial curvature was not yet available to him.

1911. Laue extended the energy-momentum tensor from the electromagnetic field to general matter systems (von Laue, 1911). Only now was the object that a relativistic theory of gravity had to couple to fully in hand.

1912. Abraham proposed a theory of gravitation with a variable speed of light. Einstein showed that it was not self-consistent (Einstein, 1912).

1912–1913. Mie derived gravitation from a theory of matter (Mie, 1913).

1912–1913. Nordström built the scalar theory. The d'Alembertian replaced the Laplacian, the trace of $T_{\mu\nu}$ replaced mass density as the source, and the theory explicitly required $m_i = m_g$ (Nordström, 1913).

1913. Einstein and Grossmann published the Entwurf (Einstein & Grossmann, 1913). It failed the general covariance Einstein himself demanded, and the perihelion calculation gave $18''$ rather than $43''$. He worked with it for two years.

1914. Einstein and Fokker recast Nordström's theory in geometric form. Its spacetime turned out to be conformally flat, which is why it predicted exactly no deflection of light.

1914. Freundlich's expedition travelled to the Crimea to measure the deflection, which at that date would have tested the value $0.83''$. War broke out, the party was interned, and no observation was made.

1915, November. Einstein rebuilt the field equations from general covariance and only then computed Mercury, obtaining $+43''$ (Einstein, 1915a;b). The perihelion was a retrodiction; the deflection of $1.75''$, twice his 1911 value, was the one prediction made in advance.

1915, November. Hilbert obtained the same field equations from a variational principle, a starting point different from Einstein's (Hilbert, 1915).

1916. Schwarzschild found the exact solution for the field outside a spherical mass, weeks after the field equations appeared (Schwarzschild, 1916). Einstein's two numbers had been obtained by approximation.

1918. Nordström, working inside general relativity, solved the coupled Einstein–Maxwell equations for a charged spherical mass, the solution now called the Reissner–Nordström metric (Nordström, 1918).

1919. The eclipse expeditions reported a deflection near $1.75''$ (Dyson et al., 1920). The measurement separated three standing values: exactly zero from the scalar theory, $0.83''$ from Einstein's 1911 equivalence-principle calculation, and $1.75''$ from the field equations.

1922. Whitehead kept a uniform Minkowski metric and built the gravitational field from retarded integrals over the sources, giving a second metric that reproduced general relativity on the classical tests (Whitehead, 1922).

1922. Abraham died without having accepted general relativity. He had a theory of his own, in which the speed of light varied with the gravitational potential.

1959. Pound and Rebka confirmed the gravitational redshift (Pound & Rebka, 1959). The value was one that Nordström's theory also gave, so the measurement discriminated nothing.

1974. Hulse and Taylor discovered a binary pulsar (Hulse & Taylor, 1975) whose orbital decay was later found to match the

rate expected from gravitational radiation.

2015. LIGO detected gravitational waves directly (Abbott et al., 2016), a hundred and nine years after Poincaré named them.

B. Mass–Energy Equivalence in a Lorentz–Poincaré Framework

The question is not whether Poincaré or Lorentz wrote Einstein's 1905 argument. They did not. Poincaré had already shown, in his discussion of Lorentz's theory and the principle of reaction, that electromagnetic energy E must be accompanied by an inertia E/c^2, treating the field as a fictitious fluid (Poincaré, 1900). Einstein later credited that paper with containing the formal considerations the result requires (Einstein, 1906). Lorentz's electron theory also treated part of mass as electromagnetic. The narrower question is this: if one keeps the Lorentz–Poincaré framework but makes its dynamics exactly Lorentz covariant, does mass–energy equivalence follow?

It does, once the total energy and momentum of a closed system are treated as a Lorentz-covariant pair. Let the system be at rest in one inertial frame. Its four-momentum is

$$P_0^\mu = (E_0/c, 0, 0, 0),$$

where E_0 is its rest energy. In a frame where the same system moves with velocity $\mathbf{v}$, a Lorentz boost gives

$$E = \gamma E_0, \qquad \mathbf{p} = \gamma E_0 \mathbf{v}/c^2.$$

For small v, this momentum must reduce to the Newtonian form $\mathbf{p} \simeq M\mathbf{v}$, where M is the inertial mass measured at low speed. Hence $M = E_0/c^2$, or

$$E_0 = Mc^2.$$

The same argument gives $\Delta E_0 = \Delta M c^2$, which is the form Einstein stated in 1905 (Einstein, 1905b).

Thus the aether interpretation by itself does not block $E = mc^2$. A Lorentz–Poincaré theory with exact Lorentz covariance and a covariant total energy-momentum law reaches the same mass–energy relation. What was missing historically was not the algebraic possibility, but the fully general energy-momentum framework for arbitrary matter systems, which was made explicit only later, especially in Laue's work (von Laue, 1911). Without those extra structural assumptions, Lorentz and Poincaré give only special electromagnetic cases, not the general equivalence.

C. AI for Mathematics vs. AI for Physics

An AI system doing mathematics can certify its own output. A formal proof can be checked step by step. The checker is software, and correctness is decidable relative to the axioms that were assumed. Generation and verification therefore live in the same box. That is what makes the formal-mathematics programme discussed in Section 3 coherent as an engineering goal (Hubert et al., 2025): an AI system can propose a proof, run the checker, and know the answer without consulting anything outside itself.

Physics has no such checker. A theory can be tested internally for Lorentz covariance, conservation, stability, and recovery of the known limit, and Nordström's theory passes all of these tests. It is self-consistent, covariant, and attractive. It explicitly requires $m_i = m_g$, and it gives the correct gravitational redshift. Every check that can be run at the desk returns a pass. It is also wrong. No further internal scrutiny cleanly separates it from the tensor theory. The clean discriminator is an observation, especially light deflection.

Simulation does not close the gap. A simulation runs the theory under test, so it returns what the theory already contains. It cannot supply an independent verdict. Automating the apparatus does not close the gap either. Execution can be handed to self-driving laboratories and robotic observatories (Boiko et al., 2023), and nothing in the argument concerns who operates the instrument. The claim is about where the information that settles the question comes from. It has to come from outside the theory being tested.

The dependence is not confined to a final validation step. The discriminating measurement usually does not exist until a theory calls for it. The experiments of Rayleigh, Brace, Morley–Miller and Trouton–Noble were designed to test consequences that Lorentz's contraction hypothesis had generated. The 1919 eclipse expedition was mounted because a theory had named a specific number. Theory and experiment are interleaved rather than sequential. That is why the loop cannot be closed once at the end, and why a human, or at minimum the world, sits inside it rather than after it.

One case has to be set aside. Systems that predict a protein structure (Jumper et al., 2021) or a material property (Merchant et al., 2023) are not being asked to find a new theory. The physics is settled and the hard part is scale, so the answer can be checked against what the accepted theory already implies, and the loop is much closer to closing inside the machine. Nothing here contradicts that work. The claim is about the other case, where the theory itself is what is at stake, and there the check has to come from outside.

The practical consequence for evaluation is that the deliverable from an AI system doing theoretical physics is not just a theory. It is a theory together with the observation that would falsify it. Scoring the pair is what the three tests in Section 5 are designed to do.