

Apodex launches TRACES benchmark to evaluate AI for scientific discovery

August 18, 2026 | News

The benchmark tests whether AI systems can navigate open-ended scientific problems, use tools, adapt to feedback and produce verifiable discoveries rather than reproduce known answers.



Apodex has launched TRACES, a benchmark designed to evaluate how artificial intelligence systems perform on open-ended scientific problems where the correct answer may not yet be known.

Unlike conventional AI benchmarks that rely on static datasets and predefined answer keys, TRACES places AI systems inside executable environments where they can observe, act, use tools, receive feedback and revise their approach while working towards a verifiable outcome.

The benchmark is intended to assess what Apodex describes as “discoverative AI” — systems designed to identify new findings from existing knowledge rather than reproduce information already contained within training data.

Depending on the scientific problem, a TRACES environment may include scientific literature, structured datasets, code execution, specialised scientific tools, simulators, folding engines and experimental feedback.

The benchmark evaluates both the final outcome and the process used to reach it.

“TRACES is a benchmark designed specifically to evaluate progress in discoverative AI. It brings together sophisticated efforts in scouting high-value real-world problems, assembling the tools and data needed to build executable environments, and developing a novel scoring system that evaluates not only outcomes but also the discovery process,” said Dr Sheng Wang, Lead Scientist at Apodex.

TRACES evaluates AI systems across six capabilities: Tools, Repair, Alternatives, Coherence, Evidence and Scope.

The Tools category assesses whether the system selects and correctly interprets external tools, while Repair evaluates its ability to identify and correct errors after receiving feedback.

Alternatives measures whether competing hypotheses are considered and revised as evidence accumulates, while Coherence evaluates whether the system maintains constraints and logical consistency across longer problem-solving sequences.

Evidence assesses whether conclusions are grounded in observations, data, experiments or citations, while Scope examines whether the system accurately defines where a conclusion applies and where it does not.

Brian Wang, AI Research Scientist at Apodex, said evaluating the process is particularly important in scientific discovery because the final answer represents only one part of a longer sequence of decisions.

“The TRACES process verification is what makes the benchmark unique. In scientific discovery, the answer is one line at the end of hundreds of judgments — what to try next, when the evidence is enough, when to abandon a hypothesis. The capability lives there, and scoring only the last line throws away almost all of it.”

TRACES combines outcome and process verification. The outcome verifier grades a submission against hidden ground truth where available, while the process verifier evaluates whether the reasoning and evidence supporting the conclusion meet the benchmark's criteria.

Evaluators match observed behaviour against written scoring descriptions, with each finding linked to specific steps in the recorded trajectory. An independent model reviews the evaluation, with disagreements triggering re-scoring and adjudication.

Apodex said the benchmark is designed to differentiate between AI systems that can reproduce established scientific knowledge and those capable of working through uncertainty, testing hypotheses and adapting to new evidence.

The company has assembled 423 high-value problems from a survey spanning 561 industries across 16 sectors as part of the broader TRACES framework.

TRACES is open to teams developing models, agent systems and solver frameworks, while researchers and organisations can also submit scientific problems to be converted into executable evaluation environments.