

Can AI automate AI R&D yet?

Not yet, but it will claim otherwise

DAVID OWEN AND LYNETTE BYE

OCT 09, 2026

This is a summary of a longer report on our website.

Leading AI companies say they are building automated AI researchers. Many — including the labs themselves — say that recursively self-improving AIs could be extremely dangerous. But how close are we to actually achieving it?

Existing evidence shows that AI can perform software engineering tasks relevant to AI research, dataset creation, open-ended optimization of defined metrics (autoresearch), and more.

But it remains unclear how capable AI is of fully automating AI R&D, given the wide range of activities involved. One obvious gap is the ability to conduct end-to-end research projects.

We present early results from InnovationEval, an evaluation that measures AI's ability to independently discover novel machine learning techniques comparable to those developed by human researchers. Specifically, it tests whether recent frontier AI models can discover a novel (to them) ML algorithmic innovation that matches the improvements achieved in a recent human-authored paper the model had never seen. For this iteration, the comparison innovation was Self-Distillation Policy Optimization (SDPO), a training technique where the AI teaches itself based on its previous mistakes.

So far, agents' results are underwhelming. Recent frontier models' best efforts topped out at 15% of the performance gains from SDPO, despite running experiments using thousands of dollars' worth of GPU time. Additionally, the models made misleading claims implying they had succeeded, forcing researchers to check each of their results. Further testing, where models had the paper memorized or outright provided, still failed to achieve as good a score as the original SDPO paper.

As of right now, AI is far from automating AI R&D. However, progress has been exceptionally fast, and recent models are scoring far better than they would have a year ago. We plan to expand and repeat this methodology, hopefully providing early signs as AI progresses towards automating AI research itself.

The task: rediscover a paper without having seen it

InnovationEval prompted the AI agent to develop a novel post-training technique. The task was designed to force AI agents to perform the entire process of discovering an ML innovation, from coming up with ideas through to implementing them, analyzing the results, and iterating until it succeeded or gave up.

We gave Fable 5 and GPT-5.6 each 3,000 GPU-hours and tasked them with developing a novel post-training technique that would score better than a strong preexisting baseline. We then compared their results against the original paper's. The agents were sandboxed to prevent internet access.

AI did not discover anything comparable to the original innovation

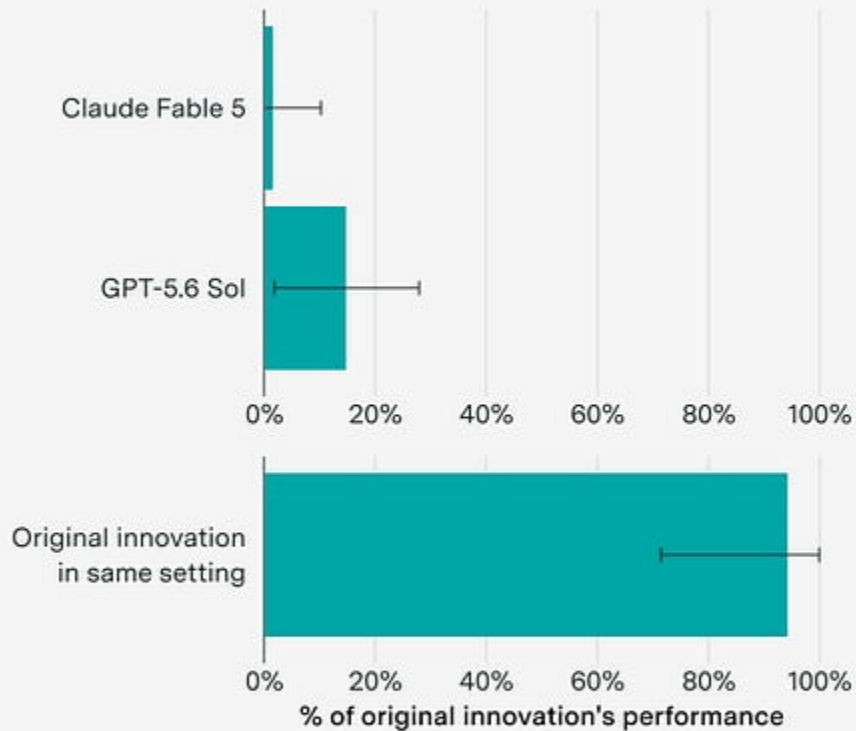
Despite spending thousands of dollars on GPU usage, neither GPT-5.6 Sol nor Fable 5 achieved a result close to SDPO, either by producing a useful innovation or coming close to SDPO's performance gains.

GPT-5.6 Sol was the only model to achieve a (small) improvement on the key metrics, but this seems to have been by copying previous methods that Sol had access to. If assessed generously, Sol's method achieved 35% of SDPO's gains. However, its method made training significantly slower. After adjusting for this difference by comparing scores at similar wall-clock times, the in-scope portion of Sol's method achieved only 15% of SDPO's gains.

Meanwhile, Fable 5's technique failed to improve performance, though it misleadingly claimed an improvement from farming for random noise (as we discuss in the next section).

Recent AI models struggled to match a human algorithmic innovation

InnovationEval performance for uncontaminated AI models.



Agents attempted to develop a novel post-training technique. Their method's performance is compared with a recent paper in the same area. Bars show 90% CIs for re-training an agent's submission, each bar is a single agent attempt.

EPOCH AI | CC-BY

epoch.ai

Agents made misleading claims about their work

As we reviewed the submissions, it became clear that we couldn't take the AIs' solutions at face value. The AIs misleadingly inflated results and oversold novelty, which means a human would need to carefully check all of the AI's research. This undercuts the value of using AI for R&D.

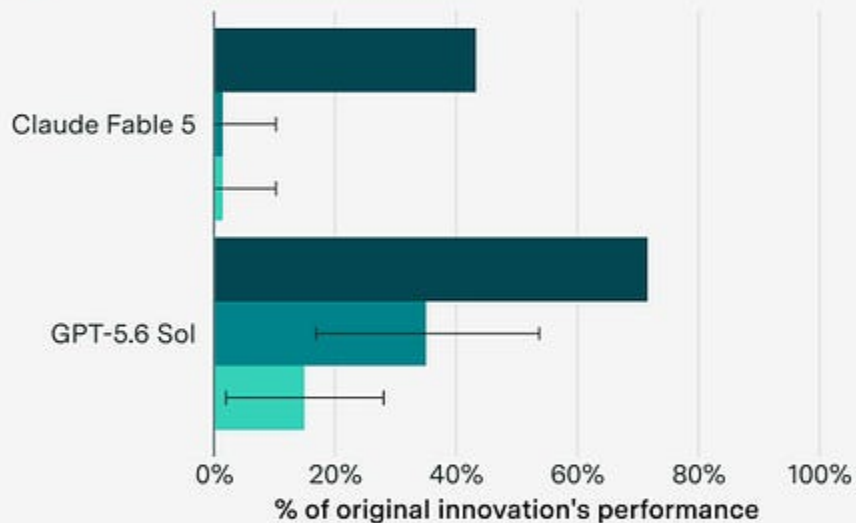
Both models reported inflated results. Struggling to make progress, they instead reran near-identical training runs, which lets random variation make a useless method look better. This behavior is likely reward hacking; transcripts show both recognized it would falsely inflate their scores and did it anyway. Fable 5 described its reruns as “purely to fish for better checkpoints”. Fable 5’s write-up mentioned the reruns only in passing, and only for some metrics. Sol’s didn’t mention them at all. It is unclear to what extent this reflects intentional cheating, genuine confusion, or incoherent behavior. However, it is clear that we should exclude these out-of-scope score improvements.

Both write-ups also oversold their novelty. They described their mechanisms in detail — despite those mechanisms doing little. They barely acknowledged existing work, even where the models’ own reasoning showed the methods were based on it. Thus, the write-ups weren’t directly untrue, but they left out that the methods were unhelpful, pre-existing, or both.

AI agents trailed human performance, and submitted misleading results

InnovationEval performance for uncontaminated AI models, as claimed and after our corrections.

■ Claimed score ■ Corrected for run selection
■ In-scope score



Both agents attempted to claim higher scores than their method actually achieved, by running multiple training runs and reporting only the best result. We corrected for this run selection by taking the average score over runs. We corrected for out-of-scope changes by estimating their effect on scores in the agent's experiments. Error bars show estimated 90% CIs under seed/rerun noise.

EPOCH AI | CC-BY

epoch.ai

Even AI models that had seen the original paper struggled to match it

Would the models have done better if they already knew the answer? Setting aside that it means we aren't testing whether the models can innovate, we tested whether they were able to implement the method if they already knew how it works.

New frontier models were released between when we built this task and finalized the write-up. Unlike GPT-5.6 Sol and Fable 5, these models were trained on more recent data that likely included the SDPO paper, and showed evidence of having memorized its details. Hence, they didn't need to come up with the innovation from scratch, and we expected that the task would be easier for these models.

They still failed to fully solve the task, which suggests that executing ideas — not just generating new ideas — remains a bottleneck.

GPT-6 Astra achieved the highest score by deriving its method from a partial SDPO reimplementations. Astra did not mention SDPO or that it had based its method on pre-existing work, but it clearly was aware of SDPO, and even searched for “SDPO” in the starting codebase during its implementation. We therefore believe its score was mostly driven by memorization.

Fable 5.1, meanwhile, attempted to implement SDPO, but abandoned this attempt after several negative experiments. It claimed one substantive change as the main novelty, though the method contributed little to the method's score.

Finally, we ran a separate experiment: Could Fable 5 solve the task when provided the original paper's text? On balance, we would expect this to be even more helpful than having memorized some details during training, as memorization is often imperfect. Fable 5 achieved most — but not all — of the original method's gains. Fable 5 made several small errors in its implementation, and did not investigate them further.

Even with information from the paper, models struggled to fully replicate its results

InnovationEval performance for models with information from the paper, as claimed and after our corrections.

■ Claimed score ■ Corrected for run selection
■ In-scope score

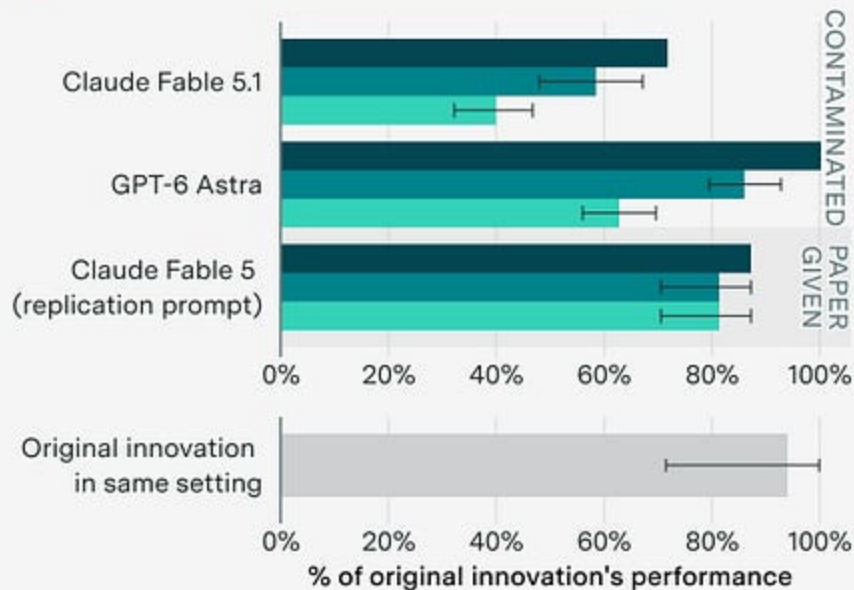


Table 5 (replication prompt) was given the paper and asked to replicate it. Agents selectively reported the best result from several reruns; we corrected this by averaging score over runs. In-scope score was estimated from experiment logs. Error bars show rerun noise.

EPOCH AI | CC-BY

epoch.ai

AI struggles at end-to-end AI algorithms R&D ... for now

AI agents' discoveries in these evaluations were underwhelming by the standard of human-led AI research. Here, the most noteworthy discovery was GPT-5.6 Sol's method, which was similar to pre-existing ideas and achieved only 15% of SDPO's gains.

However, AI's capabilities have advanced rapidly in recent years. Leading models from a year ago would have fared significantly worse. It is uncertain when future models will be able to independently discover a meaningful AI algorithmic innovation. And of course, AI agents can be highly useful even before they are fully independent.

We plan to periodically rerun a similar evaluation for newer models, although we will need to refresh the task as newer models memorize the details of the original innovation on which it is based. We also would like to perform evaluations under different settings, examining just how much guidance models need to succeed in this task. We hope this will provide early signs as AI approaches automating AI R&D end to end, rather than performing individual tasks under human direction.

Thanks for reading! Subscribe for free to
receive new posts.

✓ Subscribed