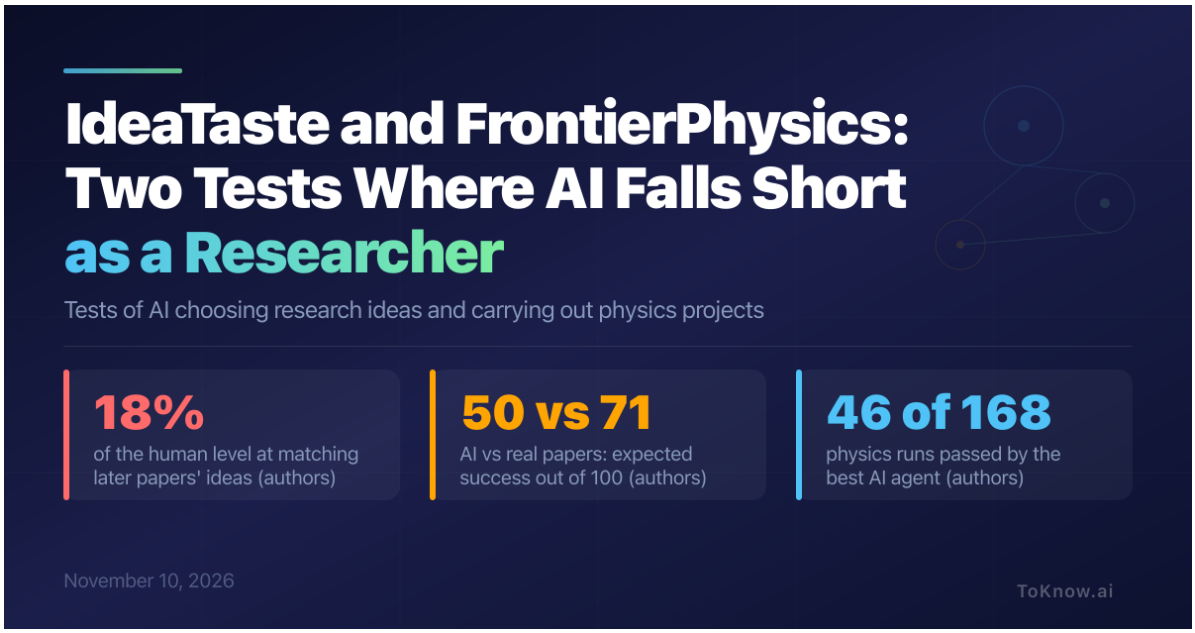
IdeaTaste and FrontierPhysics: Two Tests Where AI Falls Short as a Researcher

Kabui, Charles

2026-11-11

Read at ToKnow.ai



AI models can now prove hard maths theorems, but research also means choosing which questions to chase and seeing a project through. Two new tests measure both. In IdeaTaste, from Carnegie Mellon University, six AI models read the papers of four AI research areas up to the end of 2025 and wrote more than 90,000 research proposals. The authors compared them with the real papers published from January to August 2026. As a yardstick, they used

how much those real papers overlap with one another. The proposals reached about 26% of that level at citing the same earlier work and 18% at stating the same ideas. Eleven reviewers who did not know the source rated the proposals' chance of success at 50 out of 100, against 71 for the real papers.

FrontierPhysics asks an AI agent, a model that can search, run code and write files, to carry out a whole physics project from a short brief. Forty-two physicists wrote 56 tasks from their own research. Test scripts check the agent's numbers, and a reviewing AI grades its paper against the physicists' checklists. A run passes only if every check passes. Each of 22 models tried every task three times, and the best, Claude Opus 5.5, passed 46 of its 168 runs. In the authors' review of GPT-6 Astra's failed runs, 48% showed reasoning that was not rigorous, such as checking an answer with the same wrong assumption that produced it.

Read More: two AI labs' claimed proofs, accepted by the Lean proof checker, that the 1999 Exponential Time Hypothesis is false and a March study that trained AI on about 700,000 paper pairs to predict which paper earns more citations

Sources:

- IdeaTaste: Do LLMs Share the Research Taste of Human Researchers? (arXiv full text)
- IdeaTaste project page with the benchmark overview and live tracker (Carnegie Mellon University)
- FrontierPhysics: Evaluating Agents for End-to-End Physics Research (arXiv full text)
- FrontierPhysics: top agent clears 27.4% of physics research tasks (AI Weekly)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*