AGENT

PaperBench: Can AI Agents Actually Replicate AI Research?

Madhu Shantan

Jul 25, 2025 5 min

Remember when AI models just autocompleted your emails and churned out some Python code blocks? Those were simpler times. Now we're living in the golden age of AI agents where Cursor makes developers feel superhuman, Harvey prints enterprise money raising to $5B valuation, and Lovable boasts about that sweet $100 million ARR in <8 months. In their narrow lanes, these agents are genuinely transformative - almost feels like the start of a *tech renaissance*

Step outside those lanes? Welcome to the digital equivalent of watching a Formula 1 car attempt off-road racing.

PaperBench makes this harsh reality measurable. OpenAI's benchmark asks a deceptively simple question: Can today's best AI agents - the same ones revolutionizing software development - actually replicate cutting-edge ML research? In this blog, we'll explore how PaperBench works, what current AI capabilities reveal about research automation, and why your job as a researcher is probably safe for now.

Why This Matters: Research Isn’t Just Copy-Paste

Before you ask, “Can’t Sonnet 4 code?”—sure, it can write good code. But replicating a modern ML paper is a whole different beast. You need to:

- Parse dense academic prose (and decipher the missing bits)
- Architect a full codebase (no peeking at the authors’ GitHub)
- Run experiments, debug weird errors, and actually get the same results
- Document everything so someone else can follow along with proper context sharing

PaperBench puts AI agents through a comprehensive research challenge where they must replicate 20 spotlight/oral ICML 2024 papers entirely from scratch. Each paper comes with an author-approved rubric that breaks down the replication process into thousands of individually gradable tasks, ensuring every line of code and experimental result gets evaluated. The rules are strict: agents cannot use shortcuts, copy existing code, or take any easy paths to completion

How PaperBench Works: Rubrics, Reproduce.sh, and Ruthless Grading

Here’s the drill: the AI agent gets the paper and some clarifications (no original code allowed), then has to build a working repo from scratch. At the heart of every submission is a “reproduce.sh” script - run it, and you should get all the paper’s results,

figures, and tables. No cheating: if you hardcode outputs or fudge the process, the grading system will catch you.

Now, grading this isn't just "does it run?" Each paper has a hierarchical rubric, co-designed with the original authors, that checks everything from "did you implement the right model?" to "do the results actually match?" Rubric nodes are weighted, so nailing the main contributions counts more than getting some bonus appendix plot.

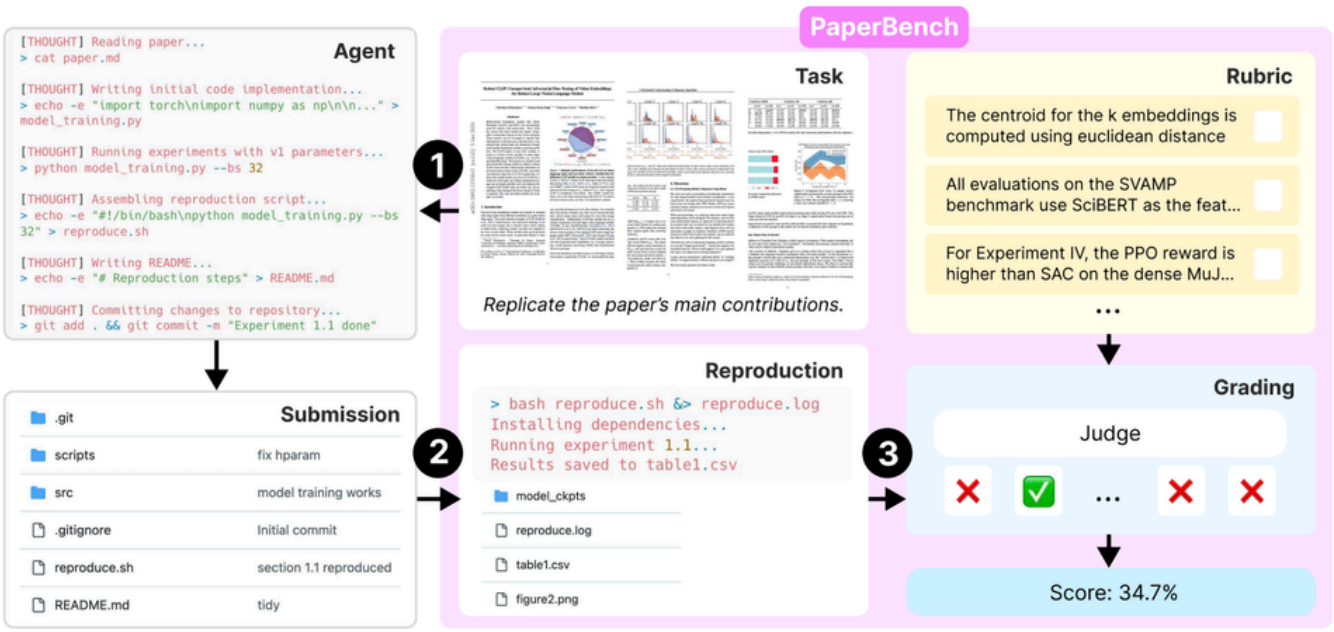


Fig: Agents create a codebase from scratch as their submission (1), which is then executed to verify result reproduction (2) and graded against the rubric by an LLM-based judge (3).

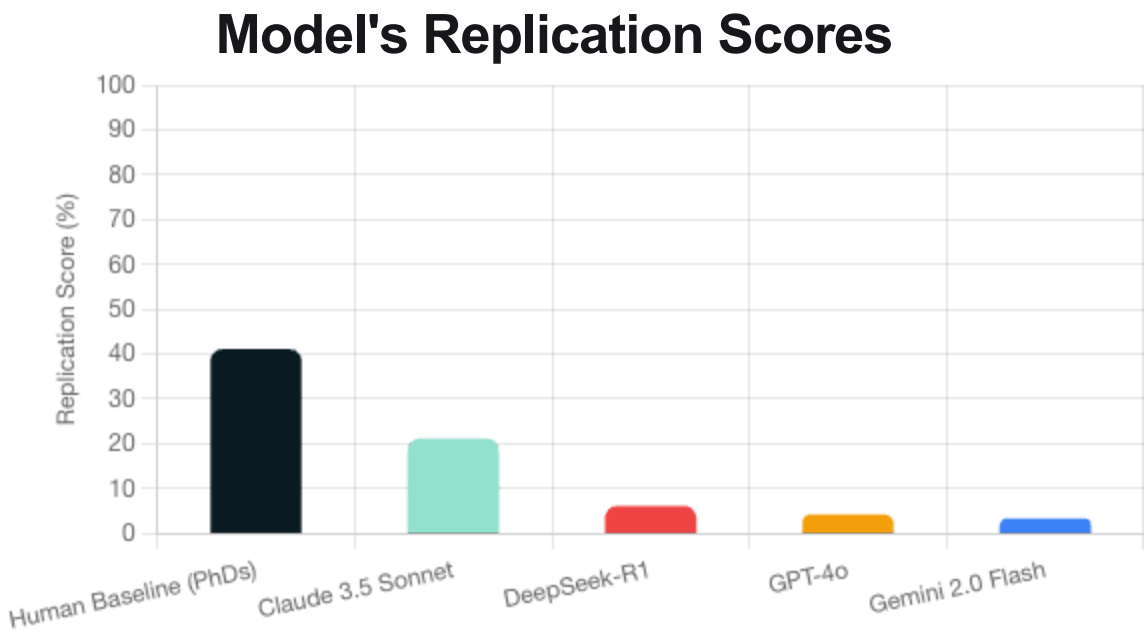
For the Evaluation of these submissions, PaperBench uses LLM-based judges using o3-mini, achieving a 0.83 F1 score against human expert judges. The main metric is the average Replication Score across all papers. It measures the weighted proportion of rubric requirements (from code implementation to result matching) that the agent successfully completes. A perfect score means the agent reproduced all key results and methods as specified in the author-approved rubric.

The Results: AI Agents Are Fast… and Flaky

The results reveal both the promise and limitations of current frontier models in research replication.

Claude 3.5 Sonnet leads with a **21.0% average replication score**, demonstrating superior persistence and strategic thinking compared to other models. OpenAI's o1 achieved **13.2%** with basic scaffolding, improving to **24.4%** with enhanced prompting that prevented early termination.

Most other frontier models, including o3-mini, GPT-4o, DeepSeek-R1, and Gemini 2.0 Flash scored **under 10%**, highlighting the extreme difficulty of the benchmark.



Average Replication Scores on PaperBench

On PaperBench Code-Dev, the lighter-weight variant that skips the execution step and assesses only code development, the agents scored much better, with o1 coming in at the front of the pack at 43.4%.

What's going wrong? Agents are great at blasting out code in the first hour, but then... they stall. They don't strategise, they don't debug, and they definitely don't sweat the details. Many agents "think" they're done after a basic implementation and just quit early. Even with tweaks to force them to keep working (the "IterativeAgent" hack), the improvements are modest.

To establish baselines, OpenAI recruited **ML PhD researchers** to attempt the same replication tasks. The results revealed that human experts achieved **41.4% average**

scores, roughly double the best AI performance.

Why Is This So Hard? (And Why Should You Care?)

Replicating research isn't just about code. It's about reading between the lines, filling in gaps, adapting to ambiguous instructions, and troubleshooting weird bugs. Humans take time to plan, experiment, and iterate - AI agents just sprint and stop (at least for now)

We're not talking about context retrieval or a deterministic programming solution. This is the real deal: can an AI agent actually do the work of an ML grad student? Also, as these agents get better, PaperBench enables us to track real progress, as we can evaluate new SOTA Agents under this benchmark!

The authors make a point to note the limitations of PaperBench - The current 20-paper dataset is limited to a relatively narrow slice of ML research. Considering the benchmarks, there's also the risk that future models might inadvertently train on published solutions, potentially inflating scores over time.

Looking Forward: The Path to AI Research Autonomy

PaperBench reveals a sobering truth: today's AI agents are impressive code generators and great at vertical fields, but terrible researchers. The gap isn't just in its alpha (raw capability), but it's in persistence, strategic thinking, and the unglamorous work of debugging and iteration that separates real research from fancy demos. Current agents treat research like a sprint when it's actually a marathon requiring patience, adaptation, and intellectual stamina.

The implications are clear: For researchers, this means your expertise in navigating ambiguity, strategic problem-solving, and scientific intuition remains irreplaceable for now. For the AI field, PaperBench provides a reality check - true research automation will require fundamental breakthroughs in how agents plan, persist, and adapt, not just better pattern matching. Nevertheless, considering there were a few model releases like the Claude 4, Gemini 2.5 Series, it is indeed worthwhile to test these models on the benchmark. Until then, AI remains a powerful tool in human hands, not an autonomous scientific collaborator.

For a deeper dive, refer to the full paper:

Paper Link

READ NEXT

AgentFold : What If AI Agents Managed Memory Like Humans Do?

NOV 11, 2025

FastLongSpeech: 30x Compression That Doesn't Murder Your Context

NOV 6, 2025

Chain-Talker: Teaching AI to Speak with Empathy

OCT 12, 2025


AICPA
SOC
aicpa.org/soc4so
GDPR
Information Security Management System
ISO
27001
Certified
HIPAA

Status