TEST KIT

OCTOBER 2026 · KIMI K3.1 PRO

CAN KIMI REPLACE YOUR AI?

Kimi K3.1 Pro is an open-weight model that Moonshot says came within a point of Claude Opus 5.5 on a coding benchmark. Benchmarks are the vendor's numbers. This kit gives you a fair, one-afternoon test on your own work, so you can decide with your own results.

01
What launched
The model, the numbers and the catch, in plain words.

02
The test
Three real tasks, one scorecard, the same rules for both AIs.

03
Decide
When to switch, when to mix, and when to wait for the weights.

01

MOONSHOT · OCT 2026

KIMI K3.1 PRO

A 3T-class open-weight model with a 1-million-token context window, out now in Kimi.ai, the Kimi apps, Kimi Work, Kimi Code and the Kimi API (kimi-k3.1-pro). A Highspeed version (kimi-k3.1-pro-highspeed) runs 2–3× faster.

65.7%
Terminal-Bench 4.0 (Moonshot-reported)

66.4%
Claude Opus 5.5 on the same test

58.2%
GPT-6 Astra on the same test

Nov 4
open weights due "by" this date

ALSO REPORTED

- On Kimi's hardest hallucination stress test (with tools), made-up claims fell from 23.4% (K3) to 8.3%.
- Strong at coding for game engines, 3D and CAD scripts, and finance work, per Moonshot.

THE CATCH

- All numbers are Moonshot's own; several use internal benchmarks.
- It trails Claude and GPT on several tests (e.g. BenchCAD 92.7 vs 96.2 / 95.9).
- Weights aren't downloadable yet.

02 THE FAIR TEST

SAME TASK, SAME RULES

Pick three tasks you really do every week. Give both AIs the exact same prompt, files and time limit. Don't help either one.

1. **A real coding or build task**
   For example: add a feature to a small app, with a plan, the code change and a test.

2. **A long-document task**
   Give it a long file or several files and ask for a specific answer with quotes. This tests the 1M-token memory.

3. **A fact-heavy task**
   Something where a wrong number would hurt: a report, a comparison, a summary of sources.

SAME PROMPT FOR BOTH · COPY

```
Task: [describe the real task]
Files: [attach the same files to both]
Rules:
1. Start with a short plan (max 5 steps).
2. Do the work. Show every change or answer clearly.
3. List anything you assumed or couldn't verify.
4. End with: "Confidence: high / medium / low" and why.
```

03 ONE SCORECARD SCORE BOTH

SCORECARD (1 = POOR, 5 = GREAT)

CHECK	KIMI K3.1 PRO · YOUR CURRENT AI
Did it finish the task?	__ / 5 · __ / 5
Was it correct? (run it / check it)	__ / 5 · __ / 5
Did it make anything up?	__ / 5 · __ / 5
How long did it take?	__ min · __ min
What did it cost?	$__ · $__
How much did you have to fix?	__ / 5 · __ / 5

ASK A THIRD AI TO JUDGE · COPY

```
Here are two answers to the same task (A and B). Don't guess which AI wrote which.
For each, list: mistakes, made-up facts, missing steps, and what you would ship.
Then say which is better for [my goal] and why, in 5 lines.
```

04 MAKE THE CALL

SWITCH, MIX OR WAIT

SWITCH IF…

- It ties or wins on 2 of 3 tasks
- It's cheaper or faster for you
- You need a huge context window

MIX IF…

- It wins on one kind of task only
- Use Highspeed for quick edits, your current AI for hard calls

WAIT IF…

- You need to run it on your own machine: weights come "by Nov 4"
- Your company needs it approved first (Enterprise is off by default)

SKIP IF…

- It made up facts on your fact-heavy task
- You'd spend more time fixing its work than you save

! **Re-test later:** models change fast. Keep your three tasks and scorecard and run them again when the open weights land.

05 TODAY WHERE TO TRY IT

1. **Kimi.ai or the Kimi app**
   Quickest way to try tasks 2 and 3 by hand.

2. **Kimi Code**
   For the coding task. Kimi Work desktop 3.1.0 or later on Windows and Apple silicon Macs.

3. **Kimi API**
   Model IDs kimi-k3.1-pro and kimi-k3.1-pro-highspeed. Zero Data Retention is available for API customers.

BEFORE YOU PASTE ANYTHING PRIVATE

- Read the provider's data policy, as you would for any AI.
- Use test copies of files, not customer data.
- Note that Moonshot deliberately limited the model's cyber-security abilities.

AI STACK SIGNAL

FOLLOW AI STACK SIGNAL
CUT THROUGH THE AI NOISE

Short, fact-checked reels on the AI tools worth your time, with the catch included.

news.aistacksignal.com/kimi-k3

SOURCES · CHECKED OCT 23, 2026

Kimi: Kimi K3.1 Pro announcement · kimi.ai

Hacker News discussion · news.ycombinator.com

Independent guide. Not affiliated with or endorsed by Moonshot AI, Anthropic, OpenAI or any company named. Benchmark figures are as reported by Moonshot on its announcement page and were not independently reproduced. Availability and the open-weights date may change. The scorecard and prompts are general suggestions.