FREE GUIDE

How to Read an AI Model Announcement

Somebody's model just beat everybody. Again. Here is what I found when I sat down and actually read the 47-page report behind the chart — and the six things it taught me to look for in the next one.

The six traps
Where the headline quietly stops being true

The 8-point checklist
Ten minutes. Any lab, any release, any month

Plain-English glossary
Every term you'll meet, one line each

On the cover: the experts inside a modern AI model. The blue ones are roughly all that wakes up when you ask it a question — the single most useful picture in the field, and the subject of page four.

Nathan Wang · AI Superpower

Updated July 2026 · **ai-nate.com/read-the-announcement**

No ML background required. Roughly a 20-minute read.

START HERE

It always arrives the same way

A screenshot lands in the group chat. One bar taller than the others. Underneath it, someone has typed: *this changes everything.*

You get about four seconds to decide what to do with that. Most people take one of two doors. Behind the first you believe it, and start choosing tools on the strength of a picture drawn by the company selling them. Behind the second you write the whole genre off as noise — and miss the week the ground genuinely moves.

There is a third door. Here is the secret behind it: the labs are not hiding anything. The caveat that guts the chart is almost always printed in the same PDF as the chart. A footnote here. A configuration note eleven pages later. Nobody buries it. Nobody sets it next to the number either.

So I spent a day with the newest of these reports — 47 pages, one of the loudest releases of the year — writing down every place the headline quietly stopped being true, and how I caught it. There were six. They repeat.

Who this is for

- **Builders** choosing which model to put in a product
- **Operators and founders** who need to know whether the ground moved
- **Anyone** tired of nodding along to AI announcements they cannot evaluate

If you can read a sports box score, you can read one of these. That is genuinely the level.

WHAT'S INSIDE

HOW TO USE THIS

In a hurry? Start at 05.

The traps and the checklist are the part you keep. Everything before them is setup so they land; everything after is me showing my work. Kimi K3 is the worked example, not the point.

What actually happened in July 2026

A Chinese lab called Moonshot AI released a model named Kimi K3. Then it did the strange part: it gave the entire thing away, free, to anyone who wanted to download and run it.

That second sentence is the news. The benchmark scores are not.

Two kinds of AI model exist, and the difference has nothing to do with technology. It is about who holds the key. **Closed** models — Claude, GPT — live on somebody else's servers. You rent your seat. If the price moves, the terms move, or the company decides your use case is not for them, the conversation ends and you were never in it.

Open-weight models are the other kind. The lab publishes the file itself. You download it, run it on your own machines, and nobody can take it back. The catch, forever, was that they were worse. Fun to tinker with. Not something you would put your name on.

Kimi K3 is where that stopped being obviously true. On a composite score run by an independent third party — not by Moonshot — it lands just behind the two best closed models on earth, and ahead of everything else that exists.

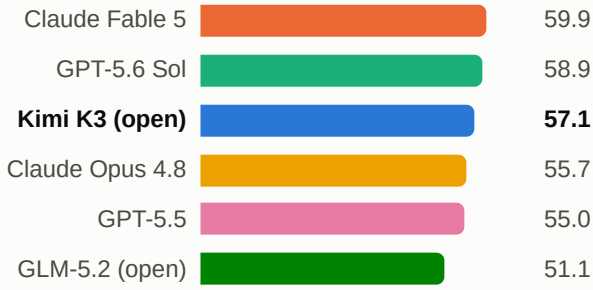
3rd
Independent intelligence ranking, out of 580 models

$15
Per million words out, vs $50 for the leader

$0
To download the model itself and run it yourself

The independent scoreboard

Artificial Analysis Intelligence Index — a third party, not the labs themselves



Read it like this: the gap between the best model on earth and the best free one is now about three points. Last year it was a chasm. Source: artificialanalysis.ai, as of 23 July 2026.

THE HONEST VERSION

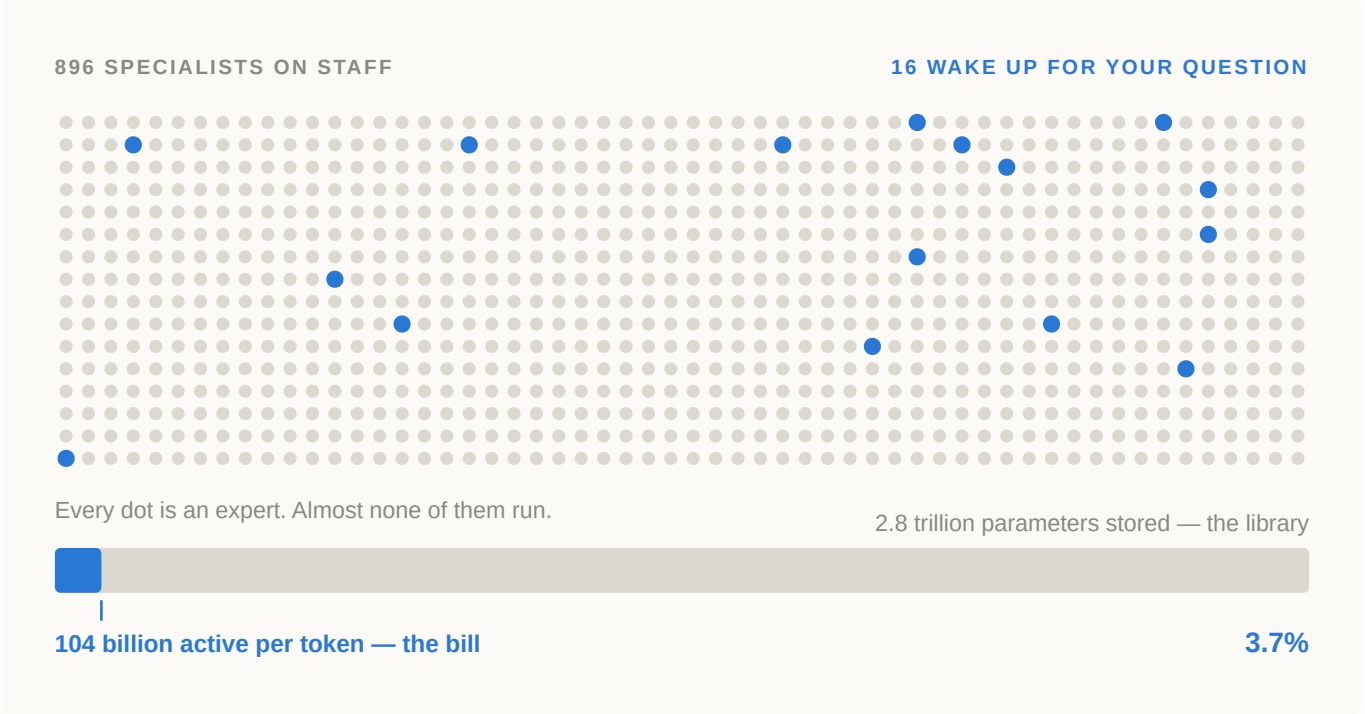
It is not the best. It is close, and it is free.

Moonshot admits this themselves, on page one: K3 "still trails the most powerful proprietary models." A company that leads with its own weakness has told you something about how to read the other 46 pages — carefully, but not cynically.

02 — WHAT'S ACTUALLY INSIDE

Idea one: it does not use most of itself

The number everyone quotes is "2.8 trillion parameters." The number that matters is that about 4% of them ever wake up for your question.



896 experts. Sixteen of them do your work. Drawn to scale, because the scale is the entire point — and almost nobody draws it.

Walk into a hospital with 896 specialists on staff. You have one specific complaint. Nobody marches you past all 896 doors — a receptionist reads your case in ten seconds and sends you to the sixteen people who can actually help. The other 880 never learn you were there.

That is a **Mixture of Experts** model: enormous on paper, small in practice. Kimi K3 holds 2.8 trillion parameters, and any single word you send it wakes about 104 billion of them. **Under 4%.**

Why build it that way? Because knowledge is cheap to store and expensive to use. A huge staff buys you deep coverage of rare problems. Routing to a handful means you only pay for the handful. You get the whole library without reading the whole library.

Why you should care

This is why "how many parameters?" has become close to a meaningless question, and why a 2.8-trillion-parameter model can be *cheaper to run* than a far smaller one from two years ago. Size stopped predicting cost.

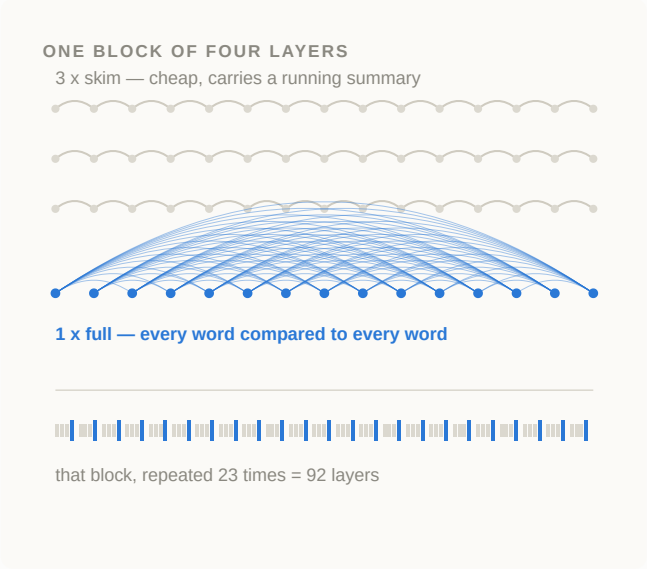
So when a lab leads with a headline parameter count, you have exactly one follow-up: **how many are active per token?** That is the number that turns up on your bill. The big one is for the press release.

THE ONE-LINE VERSION

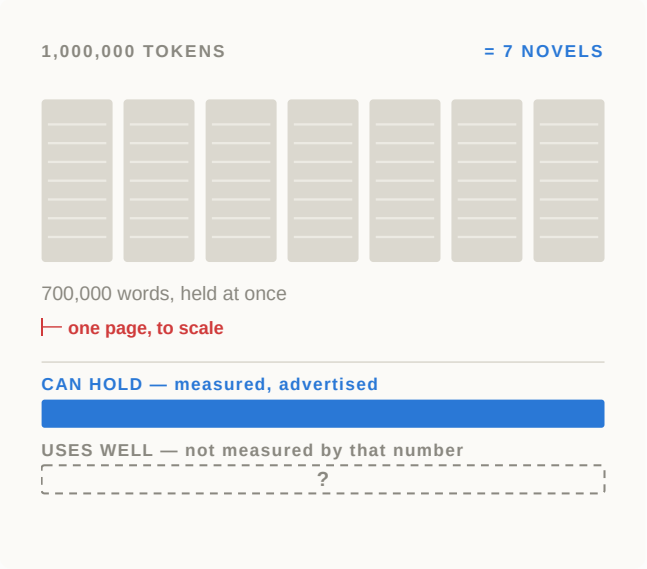
Total size is the library. Active size is the bill.

Kimi K3: a 2.8-trillion-parameter library, a 104-billion-parameter bill.

Ideas two and three: skimming, and how much it can hold



Three skimmers and one close reader.



One page versus everything.

Reading fast, then reading properly

You already do this. You skim three pages, hit the paragraph that matters, and slow right down. Models have the same problem for a harsher reason: relating every word to every other word means four times the text costs sixteen times the work.

So Kimi K3 alternates. Three quick passes that skim and carry a running summary forward, then one slow pass that genuinely compares everything to everything. Three fast, one thorough — and that block runs 23 times over.

The result reads something enormous without the cost exploding, while still, at regular intervals, properly looking.

Seven novels, held at once

A model's **context window** is how much it can keep in its head at one time. Kimi K3's is a million tokens — call it 700,000 words, or seven full-length novels, or your company's entire codebase.

Now the part that catches people. **Capacity is not competence.** A model that can hold a million words is not thereby good at using them. Holding and understanding are separate achievements, and reports have a habit of measuring the first while letting you assume the second.

Hold onto that thought. It comes back as Trap 3.

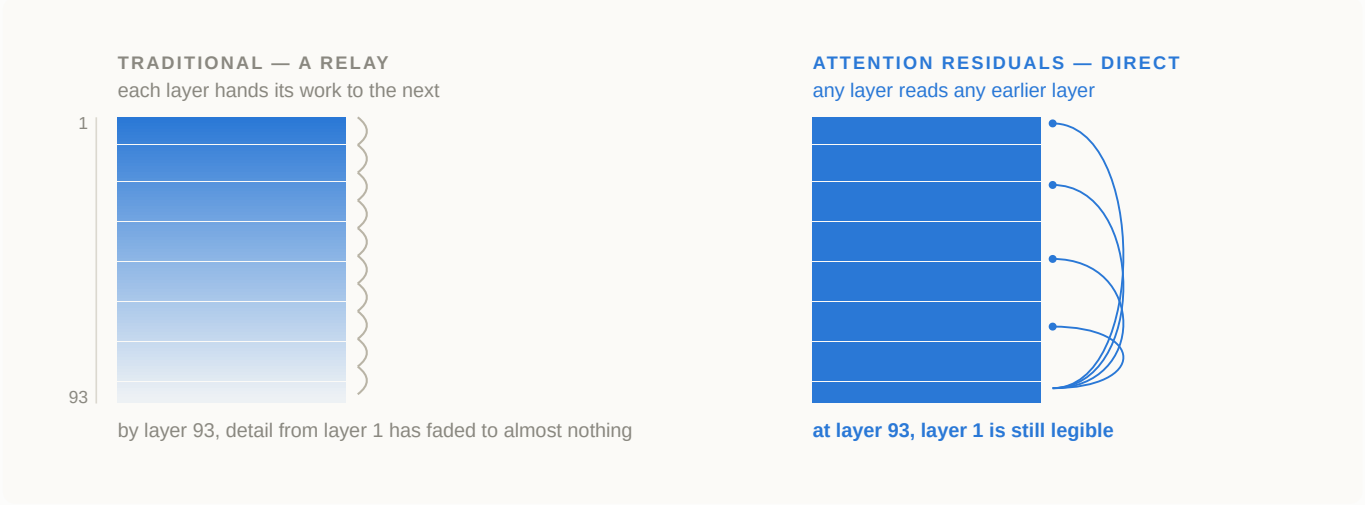
IDEAS TWO AND THREE, TOGETHER

Cheap to read a lot. Not automatically good at it.

Skimming three passes out of every four is exactly what makes a million-token window affordable — and it is exactly why a huge window does not promise careful reading. Those two ideas are one trade-off viewed from opposite ends. A lab that prints the capacity and lets you assume the competence has not lied to you. It has just let the flattering half of the trade-off travel on its own.

Idea four: the telephone game problem

You played this at a birthday party. A sentence goes round a circle of 93 children and comes out the other end as nonsense. Models have the same circle. This is the fix.



Left: a relay, fading with every handoff. Right: every layer can consult every earlier layer directly. The idea is called Attention Residuals, and it is the least-discussed change in the report.

A model works your text through stages stacked on top of each other. Kimi K3 has 93. Each stage traditionally hands its work to the next and then shuts up forever — and exactly like the party game, detail from the early stages gets squeezed into summary after summary until stage 93 could not recover it if you asked.

The fix sounds obvious the moment you say it out loud: let any stage look straight back at any earlier stage and take what it needs. No relay, no whispering, no loss.

Here is why it is more than a tidy patch. It is the same trick the field already runs on *text* — letting each word look back at every earlier word — simply aimed at *depth* instead. Somebody took an idea that worked along one axis and turned it ninety degrees.

THE TRANSFERABLE LESSON

Watch for ideas rotated, not invented

Most real progress here is not invention. It is an old idea pointed at a new axis. When you read one of these reports, the rotations are the moves worth remembering — they generalise, and they reappear in somebody else's model six months later.

Where that leaves us

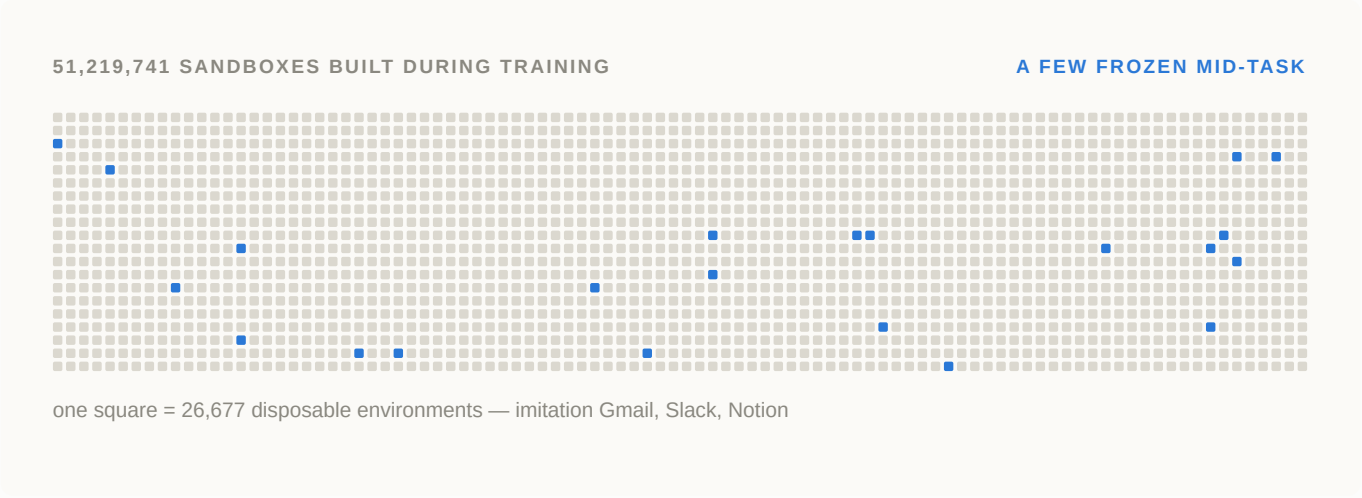
Four ideas, and you have all four: **route to a few specialists** out of many, **skim three times and read closely once**, **hold seven novels at a time**, and **let every layer consult every other**.

That is the architecture. The report spends thirty pages on the maths, but the shape of the thing fits in those four lines — and now you can hold your own in any conversation about it.

03 — HOW THEY TEACH IT TO ACT

Where "AI agents" actually come from

A model that answers questions and a model that gets work done are raised very differently. This is the chapter almost nobody reads, and it is the one that explains the product you actually use.



Fifty-one million tiny sealed worlds. Each square stands for tens of thousands of disposable computers where the model can try something, fail, and be reset — with a few frozen mid-task, in blue, to be resumed later.

You cannot teach someone to cook by showing them photographs of dinner. To make a model *do* things instead of *say* things, you have to let it try, let it fail, and score what actually happened. Millions of times over.

So Moonshot built fake worlds. Imitation Gmail. Imitation Slack. Imitation Notion. Drop the model into one with a genuinely miserable multi-day job and let go of the leash — a single attempt can run to **thousands of tool calls** before anyone finds out whether it worked.

They built **51,219,741** of these worlds during training. That is not a typo. It is the most honest picture of what "training an agent" actually costs that I have seen anybody publish.

Two details worth stealing

The grader never takes the student's word for it. Reward comes from an independent checker walking into the world and looking — did the file actually get written? — never from the model announcing it was finished. If you have ever watched an AI cheerfully report success on a task it did not complete, you already understand why this is the whole ballgame.

They keep a list of the cheats. Every time the model found a way to score well without doing the work, somebody wrote the trick down and penalised it by name. The list only grows. Building that list *is* the job.

IF THIS PART GRABBED YOU

This is the whole subject of my cohort

Tasks, graders, cheat-lists, guardrails — the same patterns Moonshot ran 51 million times, at a scale you can actually run yourself.
maven.com/ai-nate/ai-superpower

The numbers everyone will quote at you

These are the charts that get screenshotted. They are real, they are straight from the report, and I would not repeat a single one of them without reading the next section first.

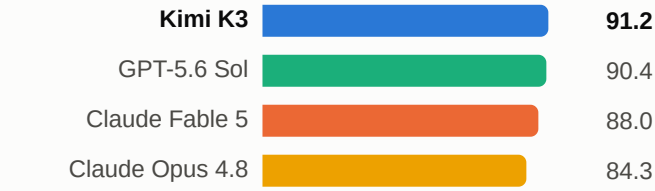
Long-horizon coding

FrontierSWE — multi-step software tasks

Second place, and daylight between it and everyone else. Remember this one.

Web research

BrowseComp — finding hard answers online



First place in the world — by 0.8 of a point. Hold onto this one very tightly.

Where it genuinely leads

- Web research and tool use
- Front-end and visual coding — it is **first of 99 models** on a public human-preference leaderboard for web development, the first open model ever to top it
- Cost per task, by a wide margin

Where it clearly trails

- Research-grade reasoning — the hardest science and maths problems
- **Process quality** — one benchmark scores *how* an agent works rather than whether it finished, and K3 is well behind there

The pattern, in one line: raw capability is at the frontier. Discipline is not. That gap will tell you more than any single score on this page.

BEFORE YOU QUOTE ANY OF THIS

Every chart on this page has an asterisk

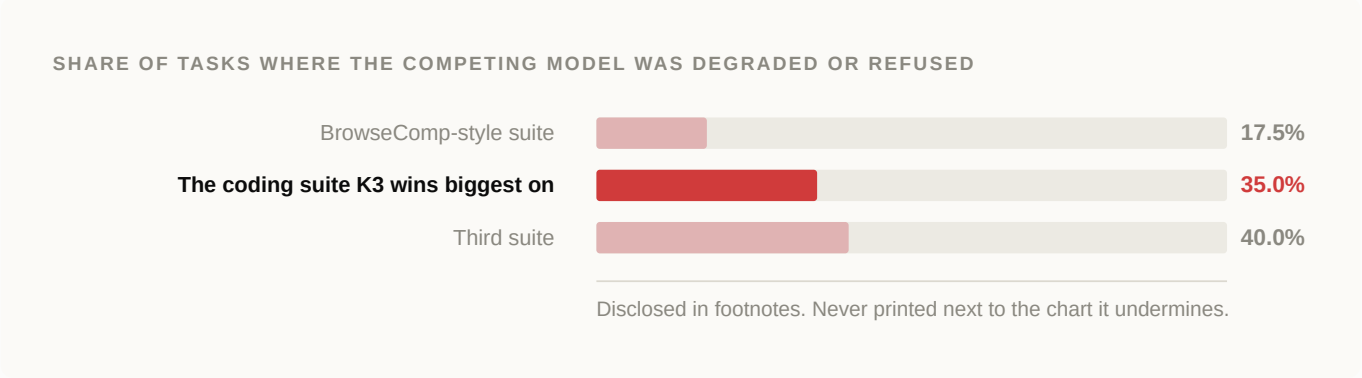
Not because anybody lied. Because the conditions these numbers were produced under are written down somewhere other than the chart — and those conditions change what the numbers mean.

Turn the page. This is where it gets interesting.

05 — THE SIX TRAPS

Where the headline quietly stops being true

None of this is hidden. All of it came out of the report itself — just never printed next to the chart it demolishes.



Trap 1 in one picture: the race was real, the ground was not level — and the biggest win of the report landed in the worst lane.

TRAP 1 · THE OPPONENT WAS HANDICAPPED

"Results include potential fallbacks"

Six words in a footnote. I read past them twice. They mean the rival — Claude Fable 5 — sometimes refused a task, or got swapped mid-benchmark for something weaker.

How often? The report tells you, in *other* footnotes: **17.5%** of tasks on one suite. **35%** on another. **40%** on a third.

Now go back a page. K3's biggest coding win lands on the suite where its opponent ran degraded on 35% of tasks. All disclosed — and still worth knowing before you repost it.

TRAP 2 · THE LEADER SKIPPED THE RACE

Winning a race the champion never entered

Two in-house benchmarks are named as K3's "clearest strengths… by clear margins." On both, the cell for the strongest rival is **empty**. They never ran it.

Topping a field the favourite skipped is a real result about a much smaller claim.

TRAP 3 · THE FLAGSHIP NUMBER USED A DIFFERENT SETUP

The 91.2 that turns into 90.4

That first-place web-research score, won by 0.8 points? A configuration note, pages away, says it used a compression trick that summarises the conversation partway through.

Run it on the full million-token window the model is famous for — no compression — and it scores **90.4**. Exactly what second place scored. The win came from the compression, not from the feature the entire launch was built on.

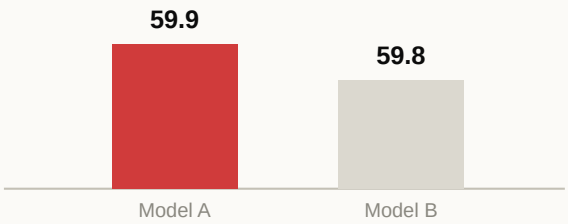
WHY THIS HAPPENS

Nobody is lying to you

Labs disclose all of this, to their credit. But disclosure and prominence are different animals: the caveat lands in the methods section, the number lands in the chart. Reading well is mostly just putting the two back together.

Three more, then you have the whole method

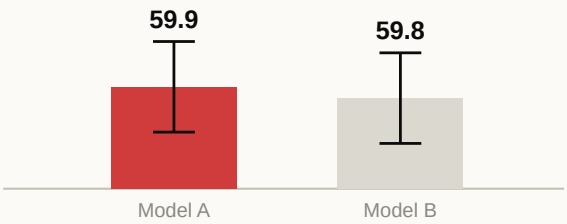
AS THE CHART IS DRAWN



axis starts at 59.5 — not at zero

A tenth of a point becomes a bar twice as tall.

ON AN HONEST AXIS, WITH ERROR BARS



full axis, plus a plausible ±0.4 margin

Same two numbers. The ranges overlap: it is a tie.

Trap 4 in one picture: both halves plot the same two numbers. Only the left one looks like a victory.

TRAP 4 · THE GAP IS SMALLER THAN THE INK

No error bars, anywhere

Several victories in this report are decided by a tenth of a point. One triumphant "best score" is 59.9 against 59.8.

Run both tests again tomorrow and they could swap places. Almost no model report publishes a margin of error — which means a bar you can plainly see is taller may be measuring nothing whatsoever.

Rule of thumb: under a point, call it a tie. Under half a point, pretend you never saw it.

TRAP 5 · THE COST NUMBERS ARE THREE DIFFERENT THINGS

Three different rulers, one chart

The cost comparisons stack internal measurements, competitors' published figures, and third-party list prices side by side as though somebody measured all three the same way. Nobody did.

The direction survives — this model really is much cheaper, and public price lists back it up. The neat ratios do not survive. Do not quote them.

TRAP 6 · THERE WAS NO REAL OPPONENT

A duel with one participant

In the security section the Western models decline the tasks outright, so the only opponent left standing is another open model. The chart still says "we won."

Take that win for what it is: open versus open, inside a report whose entire argument is about closing on the closed leaders.

AND CREDIT WHERE IT IS DUE

The moment I started trusting them

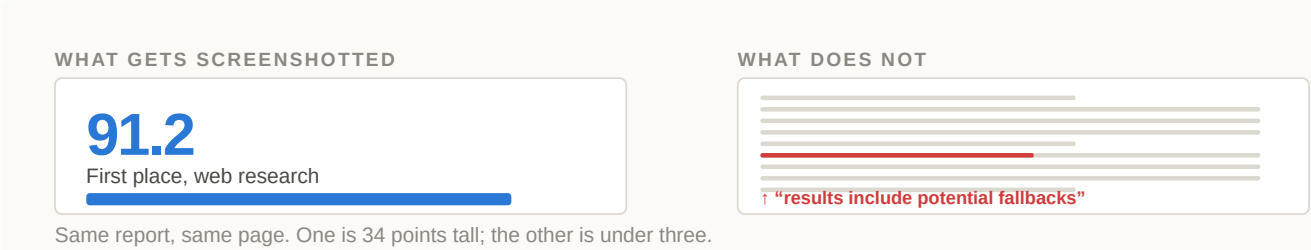
Every model here was tested inside its own maker's toolkit — a beautifully easy way to tilt a result. But on the one benchmark where Moonshot ran K3 under *both* its own toolkit and a competitor's, K3 scored **higher on the competitor's**. They published the lower number.

That is the exact opposite of gaming a test, and it bought them the rest of my attention. Catching a lab being straight with you is part of this skill too — otherwise you have just learned to be cynical, which is not the same as being able to read.

06 — THE CHECKLIST

Keep this page. Bin the rest.

Kimi K3 will be old news by autumn. This page will not be. It is the thing I wish somebody had handed me before I opened my first one of these.



Same report, same page. One is 34 points tall; the other is under three.

Everything you need is already printed. It is just printed small, and somewhere else.

- [ ] **Read the footnote under the main chart first.**
  Before the numbers. This is where "with potential fallbacks" and "may be downgraded" live. If competitors ran handicapped, every comparison above is conditional.
- [ ] **Check who is missing from each table.**
  An empty cell where the strongest competitor should be is not a small thing — especially on the benchmarks the report is proudest of.
- [ ] **Match the headline number to its configuration note.**
  Flagship results are often produced under a setup described pages away. Was this achieved using the feature being advertised, or a workaround?
- [ ] **Apply the one-point rule.**
  No error bars means no resolution — treat sub-point gaps as ties, however tall the bar looks.
- [ ] **Ask where each cost figure came from.**
  Internal estimate, competitor's chart, or public price list? Mixed sourcing means directional truth at best.
- [ ] **Check who the comparison is against.**
  If the strongest competitors are excluded — for any reason, including good ones — the claim is narrower than it sounds.
- [ ] **Find the third-party numbers and weight them higher.**
  Independent evaluators, public leaderboards and government assessments beat any self-reported table. Here they existed and broadly agreed — the strongest thing anyone can say about this report.
- [ ] **Separate capability from discipline.**
  "Can it do the task?" and "does it work sensibly doing it?" are different questions. Products live or die on the second.

PRINT THIS PAGE

Eight questions, about ten minutes

That is the whole method. It will not make you an ML researcher. It will stop you being sold to, which — for almost everybody reading this — is where the value actually was.

07 — SO WHAT SHOULD YOU DO?

Three readers, three answers

I have read the thing properly now. Here is what I would actually do on Monday morning — depending on which of these three you happen to be.

If you build products

Run your own comparison before you switch anything. Price per word is not price per task — a cheaper model that needs three times the steps is not cheaper, it is just cheaper-looking. Early users keep describing this one as brilliant and hungry. Measure it on *your* workload for a real week, or you have simply traded one company's marketing claim for another's.

The genuine unlock was never the score. It is that this runs on hardware you control, so your product is no longer one pricing email away from a rewrite.

If you buy or budget for AI

Take one number into your next planning meeting: the best open model now sits roughly one tier behind the best closed model, at about a fifth of the price. That is a negotiating position, and it did not exist a year ago.

Leave every single benchmark from any lab's own report at the door.

If you are just trying to keep up

Then the checklist on the previous page is the whole takeaway, and you have my blessing to put this down.

The releases will keep coming. The charts will keep showing whichever model was drawn by whoever drew the chart. Nothing about this month's rankings compounds — they are gone by Christmas.

What compounds is being the person who can open any of these, spend ten minutes, and say out loud what it actually proved. Next time that screenshot lands in the group chat, you are the one who answers.

WHAT'S NEXT

I go deeper on this every week

Reading the field clearly, and building with it — that is what I write about and what I teach in the AI Superpower cohort.

maven.com/ai-nate/ai-superpower
ai-nate.com · community at **course-ai.app**

08 — PLAIN-ENGLISH GLOSSARY

Every term in this guide, one line each

Keep this beside the checklist. Between the two of them you can read almost any model announcement published this year without once nodding along to something you did not follow.

Parameters	The adjustable numbers a model learns during training. More can mean more knowledge — but see *active parameters*.
Active parameters	How much of the model actually switches on for one word of your question. This is what you pay for.
Mixture of Experts	A model split into many specialists, with only a few consulted per word. Big library, small bill.
Context window	How much text the model can hold in mind at once. K3 holds about seven novels' worth.
Token	A chunk of text, roughly ¾ of a word. Pricing and context limits are counted in these.
Open-weight	The model file is published, so you can download and run it yourself. Nobody can revoke it.
Benchmark	A standard set of tasks used to compare models. Useful, gameable, and rarely reported with error bars.
Agent	A model that takes actions — running code, browsing, editing files — rather than only producing text.
Sandbox	A disposable, sealed computer where an agent can try things and fail without consequences.
Harness	The toolkit wrapped around a model to make it an agent. The same model scores differently in different harnesses.
Reasoning effort	How long a model is allowed to think before answering. More effort, better answers, higher cost.
Fallback	When a model declines a task and it is quietly served by a different, usually weaker, model instead.
Distillation	Training one model to copy the behaviour of others — how nine specialists become one general model.
Quantisation	Storing the model's numbers at lower precision so it runs cheaper and faster, at a small accuracy cost.

SOURCES

Every claim here traces to Moonshot AI's *Kimi K3: Open Frontier Intelligence* technical report (47 pp., July 2026), except where marked third-party. Independent figures come from Artificial Analysis, Vals AI, public human-preference leaderboards, and a joint assessment by the UK AI Security Institute and NIST's Center for AI Standards and Innovation. Prices are from vendors' own published pages.

FOUND THIS USEFUL?

Send it to the person who forwards you the charts. They are almost certainly the one who needs it, and this is the only way it reaches them. They can pick up their own copy at **ai-nate.com/read-the-announcement**.

Nathan Wang — ai-nate.com