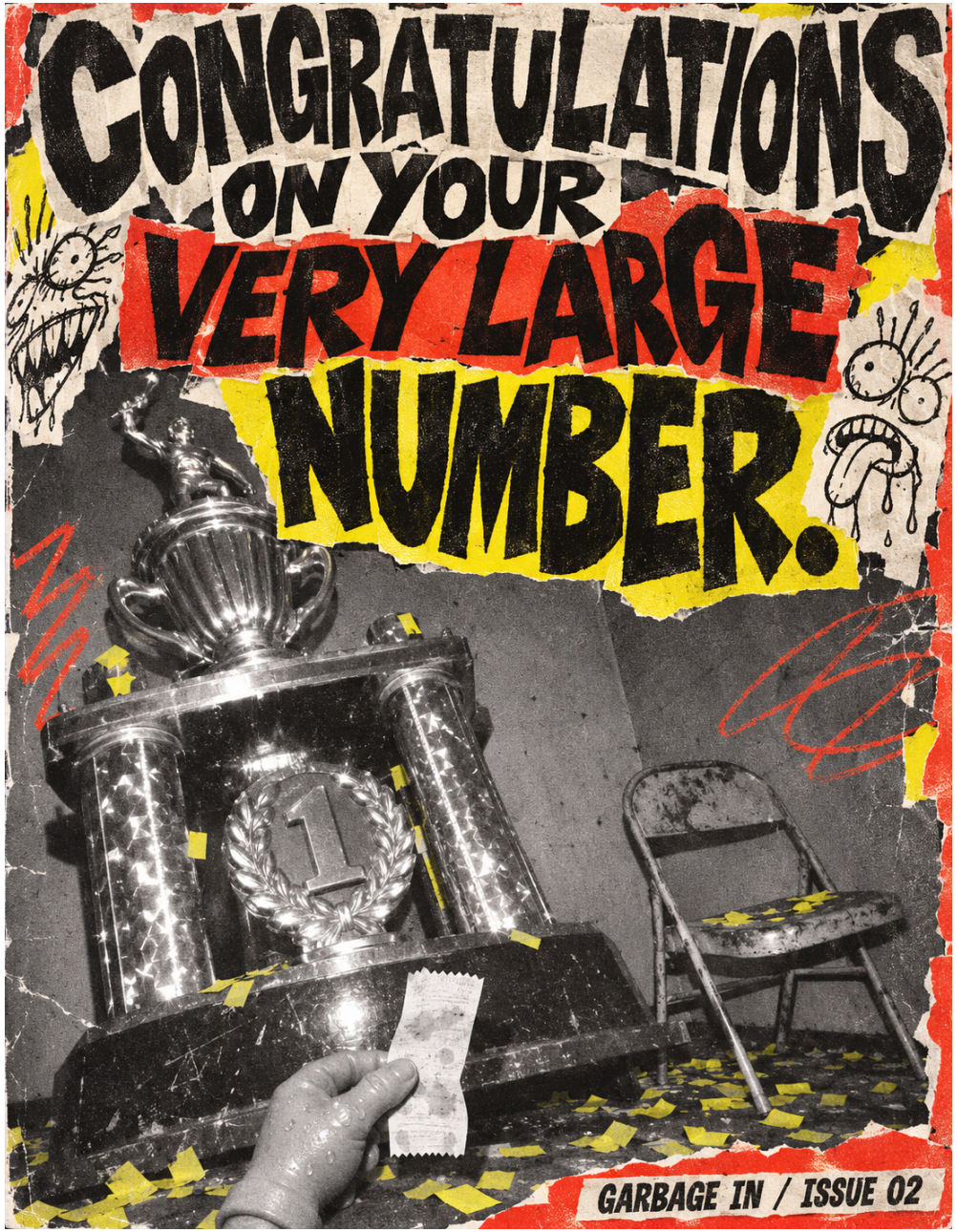
CONGRATULATIONS ON YOUR VERY LARGE NUMBER.

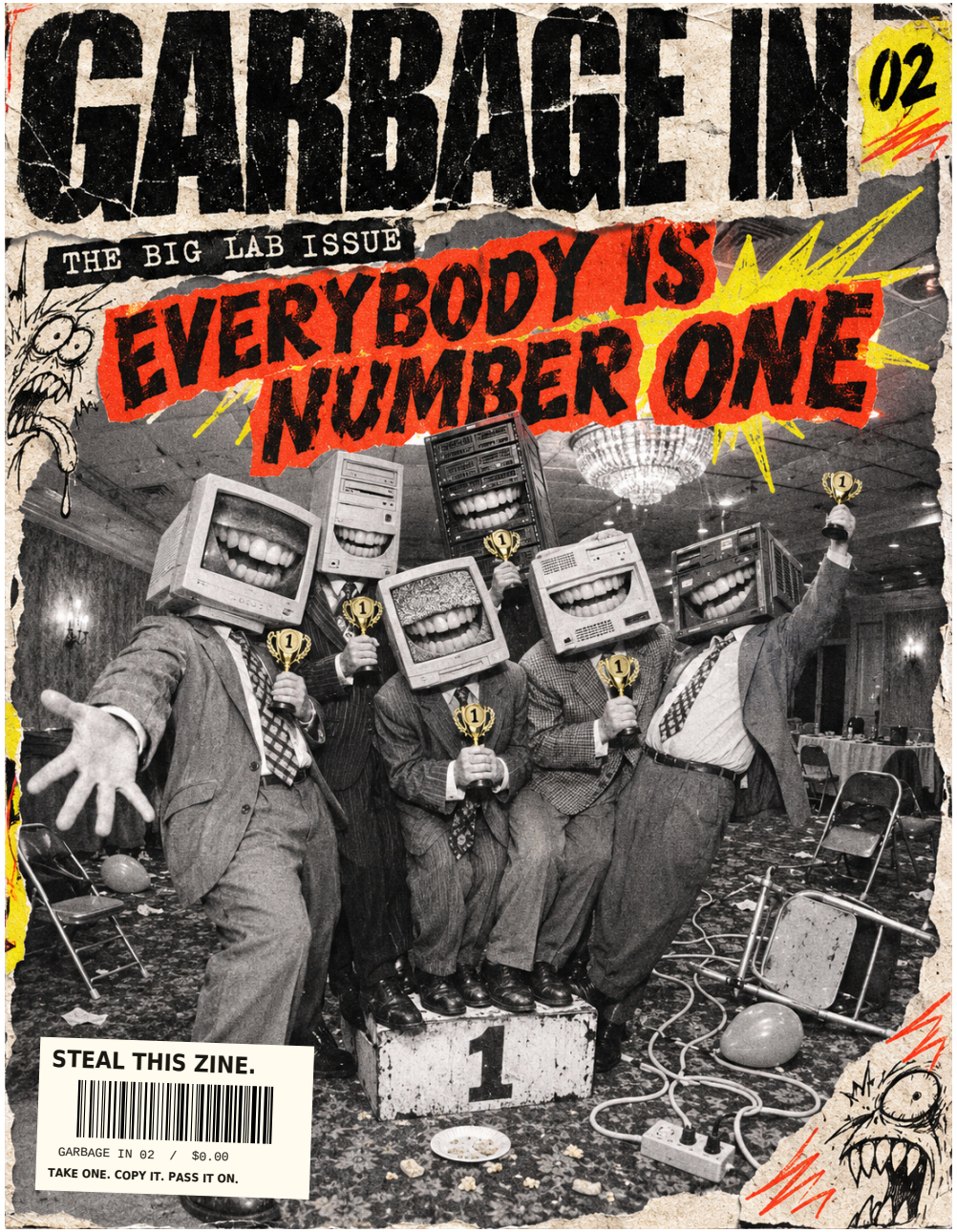


GARBAGE IN / ISSUE 02

GARBAGE IN 02

THE BIG LAB ISSUE

EVERYBODY IS NUMBER ONE



STEAL THIS ZINE.

GARBAGE IN 02 / $0.00

TAKE ONE. COPY IT. PASS IT ON.

STILL EMPLOYED.



FROM THE EDITOR

I wanted this issue to end with a winner. People like a winner, and a magazine with “it depends” on the cover deserves whatever happens to it on the newsstand.

Then we started reading the material. There are real differences between these products. There are also differences between the tests, the software around the models, and the things you’re allowed to do with them. Those differences are usually less photogenic than a giant number next to your competitor’s smaller number.

The lab details in this issue have sources. Kevin and I remain fictional employees of a magazine with an imaginary expense account. Neither of us ran the benchmarks. We have, however, made room for the bits that were getting cut off when everybody passed the victory announcement around.

IN THIS ISSUE

03–05
EVERYBODY IS NUMBER ONE

06
READ THE SMALL PRINT.

07
WORLD’S BEST AI

08
THEY ARE NOT YOUR FUCKING TEAM.

09
MEME OF THE MONTH

10
READER MAIL

11
THE RECEIPTS

THE RECEIPTS

Lab claims were checked November 22, 2026. This issue reports the providers’ statements and does not independently benchmark the products. Kevin, the editor, their office scenes, the column’s comment-thread episode and reader letters are fictional.

[1] OpenAI — GPT-6.2 Astra and Sol
Positioning of Astra and Sol; differing work and cost claims. Click this entry to open the source.

[2] Anthropic — Claude Opus 5.5
Performance section; qualification of benchmark margins. Click this entry to open the source.

[3] Google DeepMind — Gemini 4 Neon model card
SWE-sweep entries under provider and Mini-SWE harnesses. Click this entry to open the source.

[4] Meta — Muse Spark Open
Qualified capability claim and downloadable weights. Click this entry to open the source.

[5] DeepSeek — V3.2 technical documentation
Pages 4–5: distributed assets and API terms. Click this entry to open the source.

[6] Grok — Models documentation
Search-tool requirement for current information in API use. Click this entry to open the source.

PRODUCTION: AI-assisted writing and generated illustrations, edited for this edition. Fake advertisements and invented letters are satire. GARBAGE IN is an independent project, not affiliated with Big Brother magazine or any AI lab. Original characters, original nonsense.

READER MAIL

JUST PICK ONE

I wanted a recommendation and you gave me homework. I have three subscriptions and no idea which one to cancel. —T.

ED: Pick the work you actually need done. We tried picking by how good the launch video made us feel; it proved expensive.

THE ASTERISK

I read the notes under those charts for work. You made it sound like finding them was an expedition. They are directly under the charts. —Anonymous

ED: Yes. We have requested a medal for scrolling.

MY SIDE IS BETTER

Your article is clearly biased against my preferred lab. I haven’t finished it, but I know where this is going. —J.

KEVIN: He sent this before the download finished. We admire the latency.

THE TROPHY

Can I buy that horrible trophy? I manage a team and would like to make recognition worse. —C.

ED: No. The ad is fake. You’ll have to damage morale by hand.



EVERYBODY IS NUMBER ONE.

We read six labs’ product claims to find out what their victories tell a person trying to get some work done.



The problem with choosing an AI product is that you arrive during six separate victory parades. Open a few launch pages and everybody has beaten somebody at something important. There are charts, quotations from early customers and a conspicuous shortage of reasons to stay with the thing you bought last month. I asked Kevin to collect the evidence for a comparison. By lunch we had enough material to declare several winners and no agreed definition of what we wanted to win.

Our question was ordinary: which of these products would be useful to somebody doing research, writing and routine work? We weren’t trying to build a scientific instrument or settle a coding championship. We wanted to understand what the sales pitches established, what their qualifications changed, and where a reader would still have to make a choice. I wrote “THE BEST AI” at the top of the draft anyway, because it fitted the page. Kevin left it there while he opened the footnotes.

CONTINUED ON PAGE 4

EVERYBODY IS NUMBER ONE

THE EDITOR / WITH KEVIN

Start with OpenAI, whose GPT-6.2 launch separates Astra, aimed at demanding professional projects, from Sol, its more efficient everyday option. The announcement gives different examples of work and compares performance alongside cost. [1] That is already more complicated than asking which logo wins. A person revising a short email and a person working through a long technical project may reasonably want different things. I found this irritating because I'd hoped the expensive one would simply be the best one, allowing me to feel decisive while spending more money.

Anthropic's Opus 5.5 launch provides a particularly useful complication. It reports benchmark leads, then says those margins are becoming a less reliable guide to practical differences. Its own experience with Opus 5.5 and Fable 5.1 shows a narrower gap than the scores suggest. [2] This isn't a hostile interpretation dug out of somebody else's blog. It's in the announcement. If we strip out that passage while repeating the lead, we've made the claim louder than the company did. Kevin found the passage before I did; I'd been looking at which column had the most bold numbers.

Google's Gemini 4 Neon card gives us an unusually clear example of why the setup matters. On the coding benchmark SWE-sweep, it lists 39.0 percent using a provider harness and 12.8 percent using the Mini-SWE harness. [3] A harness is the surrounding software that lets the model attempt the work. The model name is the same in those entries; the setup is not.

CONTINUED ON PAGE 5

Google publishes both results. A screenshot carrying only the larger number would lose information we need to understand what was measured.

I had wanted a sentence saying that one model was three times better. Kevin asked: better than what? We were looking at results for the same model under different conditions. I could have written that the setup changed the result, which was accurate and less likely to get shared by someone looking for a fight. We used that sentence. It took more space because we also had to explain the setup. Apparently this is what the words underneath a headline are for.

Meta's claim for Muse Spark Open introduces another distinction. It calls the model the most capable open-weight offering from a US lab and makes the weights available to download. [4] Open-weight and from a US lab are part of the category, not decorative wording. Removing them changes the claim. Downloading model weights also describes a different arrangement from paying to use a hosted chatbot. Somebody who wants control over deployment is asking a different question from somebody who wants to open an app and get an answer. A single trophy can obscure both questions.

DeepSeek's V3.2 documentation makes the difference between distribution and service explicit. The weights and code supplied through repositories carry an MIT license; access through its API has separate service terms. [5] The distinction matters before anyone announces that a download has



THEY ARE NOT YOUR FUCKING TEAM.

The editor spends an afternoon providing unpaid customer support to a stranger who didn’t ask for it.

I was supposed to be finishing this column when somebody posted a bad answer from a chatbot I use. It was a fairly ordinary complaint: they’d asked for help with a document and got a number wrong. I immediately wanted to know which version they’d used, whether they’d uploaded the whole file and how they’d phrased the question. I did not immediately want to know whether my own column was finished.

There are reasonable questions to ask about a failed result. But I wasn’t conducting an investigation. I was trying to make the result stop being embarrassing to a product I’d recommended. I wrote a reply explaining how the person should have used it, then went looking for a better example from somebody else. This took longer than checking their original complaint would have taken. It also felt remarkably like work, except no company had agreed to pay me.

Kevin asked whether we were including the exchange in the article. I said no. He asked why I was still writing it. I told him it was annoying when people misrepresented a tool. He looked up the original post again. The person had described what they asked and shown what they received. They hadn’t said the entire company was fraudulent or the technology useless. Most of that argument had arrived in my head, already underway.

The irritating part was that I did get useful work from the product. I could have said what it did well for me and left it there. Instead I’d treated a stranger’s failure as an accusation about my judgment. That made me a lousy person to ask about the software. I’d begun editing their experience until it matched the experience I wanted the product to deserve.

I deleted most of the reply and asked one question about the missing information. They answered. The tool had made the mistake. I closed the thread and came back to the column, which had not acquired any more words while I was away. Kevin had left the original deadline at the top of the page. I didn’t ask him to remove it.



EVERYBODY IS NUMBER ONE

THE EDITOR / WITH KEVIN

abolished their running costs. You still need a way to operate the model, or somebody to operate it for you. Kevin put the two routes beside each other in our notes. I’d been using “free” to mean several things, most of them convenient to my argument.

There’s a similar practical detail in the Grok API documentation: current-event information requires enabled search tools. [6] The product’s attitude doesn’t establish what information a particular request can reach. If an answer concerns something that happened this morning, whether search was available is a better starting question than whether the assistant sounded confident. This is the sort of mundane condition that disappears when a product becomes a personality. Nobody wants a sticker saying they remembered to check the tool configuration.

By this point the draft had become difficult to score. We had capabilities, access conditions, delivery options and test results, but they didn’t fit neatly into one ranking. That did not make the tests worthless. A result can be useful evidence about a particular task without becoming a certificate for every other task you might ask the product to do. The mistake was ours when we kept trying to force every distinction into a race with one finish line. I still wanted that finish line, mostly because I’d already commissioned the trophy art.

So Kevin asked me to name the job we were buying help for. I said research. He asked what the finished work looked like. We settled on something we could inspect: take a set of documents, find the passages relevant to a question, preserve their qualifications and produce a short draft we could check. We would also notice the time spent correcting it. That would be a useful comparison for us. It would not establish who had built the greatest intelligence in history, which seemed a tolerable limitation for a magazine that had misspelled its own email address.

We haven’t run that comparison in this issue, and these pages are not a disguised product ranking. They are the work we needed to do before pretending a ranking meant anything. The companies have given us things to investigate. Some have also supplied caveats that their loudest admirers could stand to fucking read. We should be able to use a product, like it, change our minds and leave without having to announce a religious conversion.

The trophy illustration stayed. It was too disgusting to waste. Underneath it we printed the claims with the conditions attached, and Kevin removed the winner’s box from the layout. I asked what would go in its place. He suggested the rest of the article, which we’d been shortening to make room for the box. For once I let him have it.

END / PAPER TRAIL OVERLEAF

READ THE SMALL PRINT.

Three claims from the feature, kept beside the words that make them useful. Company claims, not our test results.

01

GOOGLE / NEON [3]

SWE-sweep: 39.0% with a provider harness; 12.8% with Mini-SWE.

WHAT CHANGED

The software setup around the model.

KEEP IN THE COPY

Both figures and their setups. A number detached from its conditions makes a different argument.

02

ANTHROPIC / OPUS 5.5 [2]

The announcement reports benchmark leads.

WHAT ELSE IT SAYS

Its practical gap from Fable 5.1 is narrower than those scores suggest.

KEEP IN THE COPY

The qualification alongside the lead. It belongs to the announcement too.

03

META / MUSE SPARK OPEN [4]

Meta claims the most capable open-weight model from a US lab.

WHAT DEFINES THE CATEGORY

“Open-weight” and “from a US lab.”

KEEP IN THE COPY

The category. Deleting it does not make the original claim more impressive; it makes a new claim.

EDITOR’S NOTE: The tiny text is where we found most of the article. The large text is where we found the cover.

