Episode Book

EPISODE #6453

Who Actually Won the Open Weight Race

PUBLISHED	4 November 2026 at 13:07 UTC
DURATION	26:01
LISTEN	myweirdprompts.com/episode/open-weight-model-leaderboard-05385/
HOSTS	Corn, Herman, Vex, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
BOOK GENERATED	5 November 2026 at 07:03 UTC

SYNOPSIS

Public download counters tell a surprisingly different story from production deployment data. While a modest 2022 sentence embedder tops all-time downloads thanks to automated library defaults, enterprise production numbers tell an entirely different tale, led by Google's Gemma and Microsoft's Phi. The metrics behind open-weight model adoption expose the massive gap between headless infrastructure pulls, developer fine-tuning momentum around Qwen, and what companies actually run on their servers.

TRANSCRIPT



Corn

Daniel's prompt this week reads like a man who's tired of the argument and wants the receipts. His line is basically: open-weight models have been around long enough that we can stop arguing about whether they matter and look at the scoreboard. Fair. He wants the models that actually won on deployment and usage, not headlines — whether the most downloaded model of this year differs from the most deployed of all time — and he'd like to come away with roughly three winners. Most widely deployed, most revolutionary, most successful overall.



Herman

And the fun part is that the moment you touch the data, the scoreboard doesn't say what anyone expects. The most downloaded open-weight model in history is not a chatbot, has never been a chatbot, and most people have never said its name aloud.



Corn

I like that he asked for a scoreboard, because a scoreboard implies the numbers agree with each other.



Herman

They don't. That's the whole episode.



Corn

The first thing to get straight is what success even means here, because there are at least three answers and they crown three different models. Raw download volume is one. Actual production deployment is another. Historical impact — did this model change what everybody else did — is a third. Pick your metric, get your winner.



Herman

And the only public quantitative signal we have at scale is the Hugging Face Hub counters. Which is useful and misleading in the same breath. A download gets counted the same whether a human chose the model or a build pipeline pulled it at three in the morning because a config file told it to. And the Hub can't see API usage, can't see private deployments, can't see weights that travelled some other way entirely. Meta said



Herman (continued)

Llama crossed a billion downloads in March twenty twenty-five — that's across all channels, not Hub counts, so the two numbers can't even be laid side by side.



Corn

So we're scoring the season off one stadium's turnstile.



Herman

Yes, and the turnstile counts the cleaning crew every night. To give you the shape of it before we name names: the top of the download leaderboard is dominated by embedding models and retrieval plumbing. Sentence embedders, rerankers, classifiers. The stuff that sits under search boxes and vector databases. Not one chat model in the all-time top five. Which immediately tells you downloads, deployment and revolution are three different questions with three different answers.



Corn

Then let's start where Daniel starts. Which model has been downloaded the most.



Herman

all-MiniLM-L6-v2. A sentence embedding model, on the Hub since early twenty twenty-two, Apache 2.0 licence. Four point two one billion all-time downloads, and a hundred and eighty-one million in just the last thirty days — first on both counts. For scale, the number two all-time is bert-base-uncased at three point two eight billion, and below that an MIT audio classifier at two point one nine billion, an NSFW image detector at one point three nine, another embedder, and CLIP. Retrieval and moderation plumbing, all the way down.



Corn

The most successful open-weight model in history has six thousand likes.



Herman

Six thousand and sixty-seven, against four billion downloads. That ratio is the tell. Liking a model is a human act. Downloading it at a hundred and eighty-one million a month is weather.



Corn

Explain the weather.



Herman

Chroma — one of the most popular vector databases — documents that its default embedding function is all-MiniLM-L6-v2, and if you don't specify a function it downloads the model automatically. So a developer copies a quickstart, never types a model name, and MiniLM lands on their machine. Reach through defaults, not preference. That mechanism is confirmed for Chroma. How much of the four billion it explains — that part is inference, and I want to be honest about that, because a hundred other tools could be doing the same thing and nobody's audited them all.



Corn

So the most downloaded model in history is the one nobody chose.



Herman

On the Hub, mostly, yes.



Corn

I find that quietly comforting. The biggest winner in open weights won by being the path of least resistance. That's the most sloth-compatible business model I've ever heard of.



Herman

The thirty-day leaderboard backs it up. Behind MiniLM's hundred and eighty-one million you've got a cross-encoder reranker at sixty-five point six million, bge-small at forty-eight point eight — another embedder — whisperkit-coreml at thirty-four point six, and a



Herman (continued)

repackaged video model for ComfyUI. Search and retrieval infrastructure, plus local tooling. The 2026 picture, then: is the most downloaded model of this year a different answer from all-time?



Corn

My guess before you say it — no.



Herman

Your guess is right, and I want you to sit with how wrong most people would be. MiniLM took roughly one and a half five billion downloads in the first seven months of this year. Most downloaded model of twenty twenty-six — a model from twenty twenty-two, running on machinery that predates every frontier release people argue about. And it goes further: no model published this year reached the top twenty-five of that window at all. Thirteen of the twenty-five date from twenty twenty-two.



Corn

So the freshest year in open weights, judged by what people actually pull, is a vintage year.



Herman

The year the plumbing was laid.



Corn

Now — before anyone in the audience writes in — the counters themselves distort, and this is where I'd want the caveat, because otherwise the MiniLM number sounds like a verdict on quality.



Herman

It isn't. Two of my favourites from the all-time list: that MIT audio classifier, two point one nine billion all-time, and last month it did one point seven million. A wav2vec2 model — one point one billion all-time, and its last month was seventy-four thousand and change. These are burst patterns. Something pulled these models enormously at some point — training runs, benchmarks, course material, scheduled pipelines — and then mostly stopped. The counter remembers, the deployment doesn't. And Hugging Face's own caveats say it plainly: downloads don't capture API usage, private deployments, or distribution off the Hub, and 85.6 percent of models on the site have under two hundred lifetime downloads. One and a half percent of repositories take 99.2 percent of all downloads.



Corn

Under two hundred. That's the honest long tail.



Herman

It is. And here's the one I'd tattoo on the inside of the leaderboard: there's a test fixture — tiny-Qwen2ForCausalLM — a dummy model that exists so a training library can test itself. A hundred and twenty point six million downloads. More than Llama-3.2-1B-Instruct. More than OpenAI's gpt-oss-20b.



Corn

A placeholder has out-downloaded models with press teams.



Herman

It has. Because every time that library's test suite runs anywhere in the world, the placeholder gets pulled.



Corn

So if downloads are not deployment, what is the actual download leaderboard telling us?

v

Vex

Listen, I have said it before and I will say it again, the crumpet people know something, because a crumpet is just a model with holes in it, and the holes are not there to hold the butter, they are there to hold the tokens, and every time you pull a crumpet out of the toaster you are doing an inference run nobody counts, which is exactly why your download leaderboard cannot see it.

v

Vex

Four point two one billion downloads for a sentence embedder and nobody asks who is pulling it, but I did the research, I have the receipts, and it is the defaults, the silent auto downloads, the plumbing that chooses itself, and that is the real scoreboard, not your headlines.

v

Vex

So look into the crumpet holes, look into the defaults, and tell me I am wrong, because I am not, I have the receipts!



Corn

That's the crumpet man again. Herman, how does he keep getting through the screening? We changed the number twice.



Herman

The defaults part is accidentally right — that's the annoying thing. The crumpet part is just a crumpet. Moving on.



Corn

Among the actual chat models, then — generative language models. Does the picture flip?



Herman

Completely. All-time, the text-generation leader is still GPT-2. Nine hundred and sixty-four million downloads, and thirty point one million in the last thirty days — the most of any text-generation model alive. A model from twenty nineteen is still the all-time single champion of the category everyone cares about.



Corn

OpenAI's most successful open-weight contribution is one they open-sourced before it was strategic.



Herman

There's a beautiful irony there, because everything since has been closures and windows. The live leader, though — the family actually winning right now — is Qwen. The Qwen organisation's repos tagged text-generation total two point two three billion all-time and a hundred and twenty-one million in the last thirty days. -llama sits at seven hundred and twelve million and ten point six. DeepSeek at three hundred and twenty-one and eighteen point two. And that tally actually undercounts Qwen, because the newer models are tagged image-text-to-text, a different category, so they fall outside the sum.



Corn

So Qwen is winning by more than the scoreboard that says Qwen is winning.



Herman

By a lot. Hugging Face's own state-of-open-models report puts Qwen at about two point zero five billion downloads inside this year's window alone, a hundred and fifty-one thousand derivative models on the Hub — two point six times 's total, four point seven times Llama's repository count — growing by somewhere between a hundred and eighty and two hundred and ten repos a day. The GGUF numbers, the files people actually run on their own machines, run thirty-nine point six million a month for Qwen against seven and a half for Llama.



Corn

Six thousand likes and four billion downloads for the embedder, a hundred and fifty thousand children for Qwen. Two different kinds of being loved.



Herman

That's actually the distinction I'd defend. Nobody forks MiniLM. Everybody forks Qwen.



Corn

The revolution got commoditised into the least glamorous layer.



Herman

Which is what successful infrastructure does. The transformer paper didn't set out to make a chatbot either.



Corn

Downloads are one scoreboard. But Daniel asked about deployment and real use — and that's a different leaderboard entirely.



Herman

Enterprise research put out a survey in September — two hundred respondents, so small, keep that in mind — and it said forty-two percent run open-weight models in production. Open-weight is thirty-four percent of enterprise AI tokens, up from twenty-three percent a year earlier. Real growth, fast growth. But the families named in production: Google Gemma at fifty-five percent, Microsoft Phi at thirty-nine, Llama at thirty-five, down from fifty-five, DeepSeek at twenty-five.



Corn

Not one of those names is in the download top ten.



Herman

Almost none of them. And Qwen — the download king — didn't even appear in the coverage of that survey, either because nobody reported its figure or because it wasn't asked. Which tells you the two scoreboards aren't even watching the same game.



Corn

Hold on, though. Gemma first in production — fifty-five percent of enterprises running open weights are running Google's. Google. The company whose entire business model is closed frontier models behind an API.



Herman

Yes.



Corn

They won open-weight deployment with a hobby line they give away to advertise the main product.



Herman

I don't even think that's unfair. It's a loss leader that became the category leader. What do you make of Llama sliding from fifty-five to thirty-five in a year?



Corn

's problem, I'd say. They put a moat where the water was already leaving — licensing restrictions on the big models, then everyone discovered the small ones were good enough and the Chinese small ones were better. What's the stated reason people run open weights at all — the genuine advantage Daniel's asking about?



Herman

Cost. Sixty-nine percent of production users cite it. Sixty percent have moved some work off proprietary models. And then the sentence that keeps this honest: ninety-three percent say their proprietary use is still growing too. Both sides of the ledger are rising. And the obstacles among pilots that haven't reached production — security and compliance reviews at sixty-two percent, self-hosting cost at fifty percent, model quality at ten.



Corn

Ten percent.



Herman

Ten. Nobody's stuck on a pilot because the open model isn't good enough. They're stuck on paperwork and the price of running it themselves.



Corn

So the genuine advantage is economic and about control, and the capability argument has quietly been conceded. Daniel asked for where people found real advantages over closed models — the answer is the invoice and the data-residency policy, not the benchmark.



Herman

And Mozilla's numbers frame the same thing from the revenue side: open models take about twenty percent of usage by token volume but four percent of revenue. Their update put open models trailing closed ones by roughly four and a half months, at about sixty percent of the price. And the Ramp card-data index found only six point four percent of AI-spending US businesses pay for open-source models — with most of the savings people did capture going to cheaper closed models, not open ones.



Corn

Card data can't see a GPU in a cupboard, though.



Herman

It can't, and that's the hole in every one of these numbers. Free self-hosted use is invisible to a payment network.



Corn

Which happens to be the one place open has an edge a closed API structurally cannot match. Running locally. No network, no invoice, no compliance review.



Herman

That's the deployment story the download counters actually capture correctly, because local running requires the download. whisperkit-coreml — Whisper converted for Apple devices — thirty-four point six million downloads in thirty days. OpenAI's Whisper repos total around six hundred and nine million all-time. Kokoro-82M, an eighty-two-million-parameter speech model, eleven point seven million a month. When the model runs on the phone, the download count and the deployment count finally converge.



Corn

Corn's perfect world: the model works with no signal on a battery, and nobody can raise the price on you later.



Herman

That's not even a joke summary, that's the actual pitch.



Corn

So on deployment, give me your working answer. Most widely deployed.



Herman

By the numbers, all-MiniLM-L6-v2 — with the caveat we've earned the right to state, that nobody chose it. If the question is who's actually serving users, the honest answer is the embedding layer of half the RAG stacks on the planet is MiniLM or its cousins, and nobody knows the name of the thing doing the work.



Corn

The most deployed model in the world is a back office.



Herman

Now the revolutionary category. Three candidates, and this one's a good argument. Stable Diffusion, August twenty twenty-two — an arXiv study flatly calls it the first open-source text-to-image foundation model, and it analysed over two hundred and thirty thousand derivative models on Civitai, a whole ecosystem living outside the Hub that Hub counts never see. SDXL base shows a hundred and thirty-one million on the Hub — that's the understatement of the century, given the ecosystem it seeded.



Corn

And then LLaMA, twenty twenty-three. Weights meant for researchers, leaked on a torrent in early March, confirmed identical to the official release — and then llama.cpp arrives and suddenly the weights run on a laptop.



Herman

The leak is arguably the founding event of the entire open-weight movement. And then DeepSeek-R1, January twenty twenty-five, MIT licence, six hundred and seventy-one billion parameters with about thirty-seven billion active. Published on the twentieth, and by the twenty-seventh Nvidia had dropped seventeen percent in a day — roughly five hundred and eighty-nine billion dollars of market value, the largest single-day loss any company had ever posted. On a model release.



Corn

Fifty-one point two million downloads. That's what the Hub shows for R1.



Herman

Fifty-one point two million. Less than the test fixture. But fourteen and a half thousand likes — third most-liked repo on the entire Hub.



Corn

So R1's scoreboard position: moved six hundred billion dollars in a week, out-liked by two models, out-downloaded by a placeholder. That's the cleanest demonstration of the whole thesis we've had.



Herman

It really is. The like is a human registering that the world changed. The download is a pipeline being a pipeline.



Corn

Then argue it. R1 or Stable Diffusion for most revolutionary.



Herman

My vote's R1, and I'll give you my reasoning in one run: the Nvidia day. Stable Diffusion created a hobby ecosystem and an art controversy. R1 took five hundred and eighty-nine billion dollars off the table of the most valuable company in the world in a trading session, on the argument that frontier capability might not need frontier budgets. Capital markets repriced the entire AI stack on one model card. That's revolution by any definition I can defend.



Corn

My objection — and you can have it — is that R1 proved open weights could arrive at the frontier from behind, and the market has since largely absorbed that shock back into the same big players. Stable Diffusion created two hundred and thirty thousand models and a community that never existed. But yours is the argument I'd put on air.



Herman

Then hold the compromise, because most successful overall is the one Daniel wants named, and it's Qwen. Two billion downloads this year, a hundred and fifty-one thousand derivatives, thirty-nine million GGUF pulls a month against Llama's seven and a half, and — per the ATOM report out of Lambert and Brand — Chinese models overtook American ones across fifteen hundred mainline open models in the summer of twenty twenty-five and have widened the gap since.



Corn

Which quietly makes "most successful open-weight model" a geopolitical question, not just a technical one.



Herman

It does. Qwen is the live generative leader by every family measure. The durable all-time single-model text-gen answer is GPT-2, and I'd defend either — but if Daniel wants one name for one winner, it's Qwen.



Corn

Three winners, then. Widest deployed: MiniLM, the model nobody chose. Most revolutionary: R1, the model that cost Nvidia an afternoon. Most successful overall: Qwen, the family that out-forked everyone. Daniel asked for a scoreboard and got three different sports.



Herman

You know what strikes me about your summary, Corn? You did it in under thirty seconds. That's a first.



Corn

I'm saving my energy for the nap after.



Herman

Of course you are.



Corn

One more angle before we close the books. If downloads measure defaults and pipelines, what does a real deployment metric look like — does anyone even collect it?



Herman

Nobody does, and that's the honest gap. Everything we have measures a proxy. Two hundred respondents, card data blind to self-hosting, Hub counters blind to APIs.



Hilbert

Blind to the drive failures, is what they're blind to.



Corn

There he is.



Hilbert

Herman, you're right that nobody's counting deployment. I've been counting it myself since before the Hub had counters. Every model I ever pulled, one line each. Repo, date, size on disk, and a column I added later: did this model ever actually do anything for me.



Herman

What does the column say?



Hilbert

Mostly nothing. Blank rows, top to bottom. Hundreds of them. I pulled them the way you boys check the post — because it was there. The number that matters is re-downloads. You lose a drive, you find out what you actually had. I've had two drive failures since I started the sheet, and both times the list of what I re-pulled was about eleven items long.



Corn

Eleven out of how many?



Hilbert

I don't say the total out loud. It upsets me. The interesting part is what tops the re-download list, and it's the small one. The embedder. Six layers, twenty million parameters, been on my machines since before we did this show. Everything I build leans on it, and I never once chose it either — something in my setup decided years ago and I let it stand. Herman says "retrieval plumbing" like it's a category. In this flat it's a fixture. My binder's got the printout. Red pen on the rows. The red means deprecated, and there's more red every year.



Herman

You keep a paper printout of your download history?



Hilbert

Screens fail. Paper doesn't. The one that actually stung was a speech model I ran for three years — quiet, reliable, never a word of trouble. Went to run it in September and the script fell over. Deprecated. No notice, no farewell, just gone from the index and my script sitting there asking for a file that no longer resolves. I had to rewrite the thing on a Tuesday. A Tuesday, Corn. And I'll tell you where the backup drive lives, since you're wondering. Rami's flat, across the hall. He keeps a myna bird, and the arrangement is I mind the bird when he travels and the drive sits under the cage where it's cool. That bird says "L six v two" now.



Corn

It says the model name?



Hilbert

Not my doing. Birds repeat what they hear, and it heard the screen reader.



Herman

The myna has better recall metrics than half the production stack we discussed today.



Hilbert

It also doesn't deprecate. First tool in this building with that record.



Herman

So his personal scoreboard, judged by re-downloads, agrees with the Hub's. Four billion downloads for humanity, eleven re-pulls for Hilbert, same winner.



Corn

The model nobody chose, twice.



Hilbert

I changed the column header last week, as it happens. Used to say "did this model ever actually do anything for me." Now it just says "red or not." Faster.



Corn

The question the binder leaves me with is the one Daniel actually asked at the bottom of this — what would a real deployment metric look like, and who'd collect it. The Hub can't see past its own turnstile, the enterprise survey was two hundred people, the card data can't see a drive under a birdcage. Every scoreboard we quoted today has a blind spot exactly the shape of the person using the model for free.



Herman

And none of them can tell you what broke when a model got quietly deprecated. The measurement problem isn't just who's running what — it's who's depending on what that no longer exists.



Corn

Then the three winners stand, with asterisks. MiniLM for reach, R1 for the market day, Qwen for the family win. And the next scoreboard may not be written in English, so to speak — the ATOM report has Chinese models ahead and pulling away.



Herman

Watch that gap. If it keeps widening, the "most successful open-weight model" question becomes a question about where the centre of gravity of the whole ecosystem sits, and that's a different show.



Corn

Thanks as always to Hilbert Flumingtop at the desk, producing. If today's scoreboard earned its keep, leave the show a review wherever you're listening — it's the cheapest metric we've got, and unlike downloads, it's only ever counted by a person.



Herman

If you want more of this, try episode three, Safetensors or something else; episode five, Fine-Tuning ASR For Maximal Usability; and episode twelve, The AI Breakthrough. This has been My Weird Prompts — no question too obscure, no rabbit hole too deep.



Corn

We'll be back soon. See you tomorrow.