Episode Book

EPISODE #5860

Who Can Actually Train an LLM From Scratch?

PUBLISHED	11 October 2026 at 10:57 UTC
DURATION	32:55
LISTEN	myweirdprompts.com/episode/training-llms-from-scratch-census-97380/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.3
SOURCES	20 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:40 UTC

SYNOPSIS

Daniel asked a simple question: how many organizations worldwide have trained a large language model completely from scratch? The answer turns out to be two questions wearing the same coat. At the frontier, the count is unverifiable — labs have stopped disclosing training compute. At the 1–70B parameter scale, the count is rising fast as sovereign initiatives in Korea, Switzerland, Brazil, Denmark, Norway, Portugal, and Iran ship from-scratch models. Meanwhile, the eight largest training runs since January 2025 are all American, and the top one is roughly 25x the compute of the largest Chinese run. This episode unpacks why "from scratch" is a harder definition than it sounds, why upcycling and continued pretraining muddy the water, and how new weight-forensics techniques are starting to make lineage auditable.

TRANSCRIPT



Corn

Quick one before we start. There's a number floating around that I want to put on the table, and then we'll take it apart for thirty minutes.



Herman

Go on.



Corn

Eighty-one. That's the count of large-scale models, above ten to the twenty-third floating point operations, that anyone has bothered to catalogue properly. Eighteen countries. Eighty-one models.



Herman

And the number of organizations that can actually make one from nothing is a much smaller figure, and nobody has ever published it.



Corn

Right. Daniel's question. He's asking something that sounds like it should have a number attached to it. How many organizations worldwide have actually trained a large language model completely from scratch. And he's strict about it. He says ab initio and he means ab initio. Randomly initialized weights, an entirely new pretraining run, the first version of a model family. Not fine-tuning. Not continued pretraining. Not distillation from an existing checkpoint. Not taking an open-weight model off the shelf and modifying it.



Herman

He's careful about that last part, and I want to flag why.



Corn

Because it's the loophole everybody uses.



Herman

It's the loophole everybody uses. He explicitly says reusing architecture is fine, reusing a tokenizer is fine, reusing techniques and datasets is fine. The distinction he cares about is whether the learned weights were inherited from an earlier model or developed through a new training run. That's the line.



Corn

And then he asks two things. First, as of now, how many organizations have demonstrated they can do this, and are there any credible estimates or surveys that actually answer it. Second, he wants the concrete examples. Who trained what, when, how big, and how do we know the weights were random and not inherited.



Herman

Particularly outside the famous American labs.



Corn

Particularly outside. Chinese companies, independent research shops, universities, national AI initiatives. And the broader question underneath all of it, which is whether the apparent diversity of the model ecosystem is disguising a much smaller group of organizations that can create foundational models independently. Is that concentration going up or coming down.



Herman

The answer is yes.



Corn

Both.



Herman

Both, and that's not a dodge, that's the finding. But let's do this properly.



Corn

Define the term first.



Herman

Define the term first, because the whole episode hinges on it. From scratch means randomly initialized weights and a new pretraining pass. That's it. That's the whole test. A model that started as a Llama checkpoint and got continued-pretrained on Korean text for a few billion tokens is not from scratch. A model that was distilled from GPT-class outputs is not from scratch, even if every other part of the pipeline was built in-house. A model that was shrunk from a larger sibling through pruning is not from scratch.



Corn

And a model that was upcycled, which is the new word I've been seeing everywhere and which I want us to get to.



Herman

Upcycling is taking an existing model's weights and expanding the architecture around them. You keep the learned weights, you grow the network, you train the new parts. LG AI Research did exactly this with K-EXAONE 2.0 and said so explicitly in the technical report. Their words were, rather than training from scratch, we upcycle K-EXAONE and expand its architecture.



Corn

That's a company volunteering that it did not do the thing.



Herman

It's a company volunteering that it did not do the thing, in a technical document, in August. That's a remarkable sentence to find in a release note, because the marketing around almost every model launch is designed to imply the opposite.



Corn

So what survives the definition. What counts.



Herman

What survives is a model where you scratched the weights and trained them up. You can borrow the architecture, you can borrow the tokenizer, you can borrow the recipe, you can train on the same public web corpus as everyone else. None of that disqualifies you. The question is purely whether the numbers inside the model started as noise and became structure through a training run you paid for.



Corn

Which makes the accounting question harder, not easier, because a model can look brand new and be a derivative.



Herman

And it can look derivative and be brand new. That's the part that trips people up. The architecture is not the tell. The name is not the tell. The press release is definitely not the tell.



Corn

So there are two problems stacked on top of each other. First, how many. Second, how do you know.



Herman

And the second is the one that makes the first unanswerable, which is where I want to spend the bulk of this.



Corn

Let's start with the census that doesn't exist.



Herman

There is no authoritative global count of organizations that have completed a from-scratch pretraining run. Not a disputed one, not an out-of-date one, none. The closest thing anyone has is Epoch AI's large-scale model dataset, which tracks models above ten to the twenty-third floating point operations. As of April 2024 that dataset had eighty-one models across eighteen countries. Forty-three from the United States. Nineteen from China. Six from the United Kingdom. Seventy-one of those came from industry, two from academia, two from government institutions.



Corn

Two from academia. Across the entire frontier.



Herman

Two from academia in that dataset, yes. And Epoch, to their credit, was explicit about the limit of what they were doing. They said that determining the number or identity of organizations capable of creating these models would be a useful endeavor, but was beyond the scope of their process.



Corn

Which is a very polite way of saying we counted the models, not the kitchens.



Herman

That's exactly what it is. They counted outputs. They never claimed to count capability. And everybody since then has been quoting the model count as if it were a capability count.



Corn

So what happens when you try to count capability using something looser.



Herman

You get nonsense, in both directions. Epoch's broader data now tracks seven hundred and twenty-seven models released between January 2025 and September 2026, from three hundred and sixty-five American organizations, two hundred and forty-eight Chinese, thirty-two French, twenty-three Korean, nineteen Canadian. But that's releases. Releases include fine-tunes. Releases include merges. Releases include a weekend project that took a Qwen checkpoint and taught it to speak in rhyme.



Corn

Three hundred and sixty-five American organizations releasing models.



Herman

And I would bet most of them have never initialized a weight matrix in their lives. Then you go the other direction and you get TrendFeedr, which counts nine hundred and fifty-nine companies working on foundation models with thirty-eight point two billion dollars in funding. Working on. That word is doing an enormous amount of labour in that sentence.



Corn

A company with a landing page and a seed round is working on foundation models.



Herman

So you have two proxies, one undercounts capability and one overcounts it wildly, and neither of them is a census. It's the gap between them where the real answer lives, and nobody has filled it in.



Corn

Liquid AI made a claim about this in May.



Herman

They did, and it's the most direct public statement I've found from anyone actually in the business. Their line was, there are only a few companies worldwide that train their own foundation models from scratch. And then they listed the pipeline to make the point concrete. Model research, pretraining, posttraining, deployment. Four stages. They're saying most of the industry participates in the last two.



Corn

Posttraining and deployment.



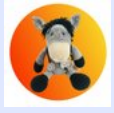
Herman

Most of the industry does posttraining and deployment. Which is not a criticism, by the way. Posttraining is where a huge amount of the useful work happens. But it's a different skill from standing up a pretraining run.



Corn

And the count changes enormously depending on the scale you're asking about, which Daniel flagged and which I think is the sharpest part of his question.



Herman

It's two completely different questions wearing the same coat. At frontier scale, above ten to the twenty-fifth floating point operations, the number is not small, it's unverifiable. We can't count it, because the frontier labs have stopped telling us how much compute went



Herman (continued)

into the training run. Claude Opus 5, GPT-6 Astra, there is no public compute estimate for them. None.



Herman

At the very top the count is unknown and probably lower than people assume, because the barrier there is capital and interconnect and power, and those are getting harder, not easier. But at the smaller scale, the one to seventy billion parameter range, the count is rising fast. That's where the sovereign and national initiatives live. Korea, Switzerland, Brazil, Denmark, Norway, Portugal, Iran. All of them have produced from-scratch models in that band in the last year or so.



Corn

So the answer to Daniel splits cleanly.



Herman

It splits cleanly, and this is the bifurcation I want to keep coming back to. The count of from-scratch models is going up. The concentration of the compute frontier is also going up, in a different direction, toward a handful of American labs. Both things are true at once and most coverage picks one and pretends the other doesn't exist.



Corn

Give me the number that makes the second half real.



Herman

The eight largest training runs since January 2025 are all American. Every one. The largest is xAI's Grok 4, at roughly five times ten to the twenty-sixth floating point operations. The largest Chinese run in that window is Moonshot's Kimi K3 from July, at about two times ten to the twenty-fifth. That's roughly a factor of twenty-five.



Corn

Twenty-five times.



Herman

Twenty-five times the compute in the top American run versus the top Chinese run. And China is not a weak player. China released eighty-two large-scale models in 2025 against sixty-six American. China released twenty-eight open-weight notable models against eleven from the US. China's share of Hugging Face downloads climbed to seventeen point one percent in the year to August, up from six point three percent all-time, and DeepSeek and Qwen alone account for fourteen points of that.



Corn

So China is winning on volume and open weights and losing by a factor of twenty-five on the single biggest training runs.



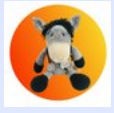
Herman

Which is a interesting strategic position, and it's not the one either side's partisans describe.



Corn

Now the harder problem. The evidence.



Herman

How do you prove a model was trained from scratch.



Corn

Because the marketing will always say it was.



Herman

The marketing will always say it was, and there was no mechanism to check until very recently, and the mechanism that now exists is the most interesting technical development in this whole story. Let me do the governance side first, then the forensics.



Corn

Governance first.



Herman

The Foundation Model Transparency Index added an indicator in its 2025 edition called Model dependencies. It asks developers to disclose the models a model is derived from. That means teacher models used for distillation. It's an attempt to force the disclosure of lineage.



Corn

And the effect.



Herman

Transparency fell. The average score went from fifty-eight out of a hundred in 2024 to forty out of a hundred in 2025. Training data and compute are the two most opaque areas, and compute is precisely the number you'd need to place a model on the scale ladder. IBM scored highest ever recorded at ninety-five. xAI and Midjourney were joint lowest at fourteen.



Corn

Fourteen.



Herman

Fourteen out of a hundred.



Corn

So the index got sharper and the industry got quieter.



Herman

The index got sharper and the industry got quieter, and I don't think that's a coincidence. The moment you start asking specifically about dependencies is the moment the disclosure drops.



Corn

Which is an admission in itself.



Herman

It's an admission in itself. Now the forensics, because this is the bit that changes the game. There's a method called AWM, weight-matrix fingerprinting, no training required. It looks at the internal weight matrices of a suspect model and determines whether it was trained from scratch or derived from an existing base model.



Corn

No training required means what exactly.



Herman

It means you don't need to know anything about the training run. You don't need the data, you don't need the compute logs, you don't need the developer to cooperate. You take the model and you look at its insides.



Corn

And it holds up against the obvious evasions.



Herman

It holds up against supervised fine-tuning, against continued pretraining, against reinforcement learning, against multimodal extension, against pruning, and against upcycling. All of those transformations preserve enough of the parent's fingerprint to be detectable.



Corn

Upcycling included.



Herman

They tested it on sixty positive pairs and ninety negative pairs and got perfect classification, and it runs in thirty seconds on a single RTX 3090. Thirty seconds on a consumer gaming card.



Corn

That's the piece that should worry people who've been vague about lineage.



Herman

That's the part. A research lab could always run fingerprinting. What AWM does is put it in reach of anyone with a spare graphics card, and it turns the question of provenance from a trust exercise into a measurement.



Corn

And the enforcement story already exists in the real world.



Herman

It does. Korea's sovereign AI contest had an explicit rule that the models had to be trained from scratch, and Naver Cloud was disqualified because its vision encoder used locked weights from Alibaba's Qwen 2.5-VL. Here's the thing about that disqualification. Naver Cloud is a serious company. This isn't a startup that got caught cutting corners. This is a national champion, in a well-funded government programme, with hundreds of



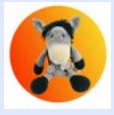
Herman (continued)

GPUs, and they either couldn't or didn't build the vision encoder from scratch.



Corn

And got caught because the rule existed.



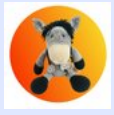
Herman

Which tells you something about how many organizations would pass that rule if anyone applied it.



Corn

It also tells you the rule is expensive to obey.



Herman

The rule is expensive to obey, and the incentive to quietly not obey it is enormous, because a vision encoder that works is worth more to a product than a purity certificate.



Corn

And LG went the other way and just said it publicly.



Herman

LG went the other way and said it publicly. K-EXAONE 2.0 is a seven hundred and fifty billion parameter mixture-of-experts, and the report states plainly that it upcycles rather than pretrains. That's the industry shifting underneath the marketing. The release cycles are getting faster, the models are getting bigger, and the actual training method is increasingly derivation from something that already exists.



Corn

Upcycling as the new normal.



Herman

Upcycling as the new normal, and the market rewards it, because a bigger model next quarter beats a purer model next year, every time.



Corn

Okay. So we have the counting problem and we have the verifying problem. Let's spend the second half on who's actually done it, because that's where Daniel wanted the specifics.



Herman

Where do you want to start.



Corn

Korea. It's the most aggressive national programme and it has a documented scoreboard, which almost nobody else has.



Herman

Korea is the best-case study in the world for this question, and I'll tell you why. The Ministry of Science and ICT selected five elite teams in August 2025 and gave each of them between five hundred and twelve and a thousand and twenty-four GPUs. Naver Cloud, Upstage, SK Telecom, NC AI, LG AI Research. The goal was explicitly to build Korean foundation models trained from scratch.



Corn

And then they ran it as an actual competition with eliminations.



Herman

They ran it as a competition with eliminations, and the phase two scores were published in August. SK Telecom seventy point six. Upstage sixty-nine point nine. LG AI Research sixty-nine point zero. Motif sixty-five point eight, and Motif was cut. Those four numbers are within about five points of each other across four multi-hundred-billion-parameter model programmes built by four different companies in one country in one year.



Corn

That's a real spread of capability in a small market.



Herman

That's a real industrial base, is what that is. And then all five of the surviving teams released models at the end of December 2025. HyperCLOVA X SEED 32B Think, A.X K1, VAETKI, K-EXAONE, Solar Open 100B. Five from-scratch entries in one month from one country.



Corn

SK Telecom's the one with the documentation.



Herman

SK Telecom's the one with the documentation, and it's thorough. A.X K1 was five hundred and nineteen billion total parameters, thirty-three billion active. A.X K2 is six hundred and eighty-eight billion total, thirty-three billion active, trained on about eight point five trillion tokens, two hundred and fifty-six thousand token context, Apache 2.0 licence. And the report says trained from scratch, in the paper, in the repo, in the model card. A.X K2 was also trained natively in FP8.



Corn

Natively in eight-bit floating point from the start.



Herman

From the start, not converted afterward. That's a meaningful engineering claim on its own.



Corn

So Korea has at least two, arguably five, from-scratch families in the hundreds-of-billions of parameters.



Herman

At least two with the documentation I'd want to see, and five with a government scoreboard behind them. Which is more from-scratch pretraining capability than most countries have ever demonstrated.



Corn

China next, because the interesting thing there is the mix.



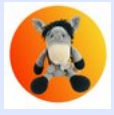
Herman

The mix is fascinating. Baidu's ERNIE 5.0, from February, is a trillion-parameter unified autoregressive model, and the technical report says all modalities were trained from scratch under a unified objective. Every modality, one objective, one training run. That's a serious claim at trillion scale.



Corn

And Ant Group.



Herman

Ant Group's Robbyant arm put out LingBot-VA 2.0 in July, a robot foundation model pretrained from scratch. Robotics foundation models are a completely different design problem from text, and they're doing the pretraining themselves.



Corn

So China's got the full stack from a trillion-parameter general model down to a robotics model.



Herman

And then the smaller sovereign initiatives, which is where the count is really rising. Switzerland's Apertus, from ETH Zurich and EPFL, released September 2025. The seventy billion version was trained on the Alps supercomputer, and the training data and methods are fully disclosed. Which is unusual.



Corn

Fully disclosed meaning you could rerun it.



Herman

Fully disclosed meaning someone could, in principle, reconstruct the whole pipeline. That's rarer than it should be.



Corn

Brazil.



Herman

Manacá-1B, from August, one point seven two billion parameters, trained from scratch for Brazilian Portuguese with a fully reproducible pipeline. Denmark's DFM Mimir v1, also August, a one-billion-parameter model on the HRM architecture, trained from scratch on permissible data, and it's state of the art for Danish. Norway's NorwAI models are either pretrained from scratch or continually pretrained on twenty-five to eighty-eight billion tokens, depending on which model. Portugal's AMALIA is a publicly funded nine-billion-parameter model. Iran's IHUBERT is a Persian RoBERTa-base encoder, a hundred and twenty-five million parameters, trained from scratch.



Corn

That last one's worth pausing on.



Herman

It is, and I'd rather state it than editorialise. A hundred and twenty-five million parameters is tiny by any current standard. It's a masked-language-model-sized encoder. But somebody in Iran built one from random initialization for Persian, and that's the point. The floor for doing this has fallen far enough that a research group in a heavily sanctioned country can produce a from-scratch model for its own language.



Corn

The floor falling is the good news half of the bifurcation.



Herman

The floor falling is the good news half. Now the American independents, because Daniel asked for examples outside the famous labs, and these are the ones I find impressive.



Corn

Arcee.



Herman

Arcee AI's Trinity family. First from-scratch pretraining for the company, mixture-of-experts, roughly twenty million dollars in total development cost, and a team of about thirty people.



Corn

Thirty people. Twenty million.



Herman

That's the number that should reframe the whole conversation. Twenty million dollars and thirty people gets you from-scratch pretraining capability in 2026. That is the price of a mid-sized commercial building.



Corn

Which is not nothing.



Herman

Which is not nothing, and the compute rental alone would eat most of it. But five years ago that number was in the hundreds of millions and required a team ten times the size.



Corn

Poolside.



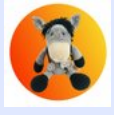
Herman

Poolside's Laguna M.1 at two hundred and twenty-five point eight billion parameters, and XS.2 at thirty-three point four billion, both released in May and both described as trained from scratch end-to-end. And Linum-V2, which is my favourite entry on this list, a two-billion-parameter text-to-video model built from scratch by two brothers.



Corn

Two brothers.



Herman

Text to video, from scratch, at two billion parameters. There are large organisations that couldn't put that together.



Corn

Okay. Concentration. Daniel's last question. Is it going up or down.



Herman

The count is going up. That's unambiguous. More organisations have from-scratch capability today than at any point in the industry's history, because the floor keeps dropping and the sovereign programmes keep funding.



Corn

And the frontier.



Herman

And the frontier is concentrating harder than ever. The top eight training runs since January 2025 are American. The gap between the biggest American run and the biggest Chinese one is twenty-five to one. Stanford's AI Index for 2026 counted ninety-three notable models from industry in 2025 and two from academia.



Corn

Two.



Herman

Two from academia. Over the past decade Tsinghua and Stanford are level at twenty-six each and Carnegie Mellon has twenty-five, but that's the decade, not the year. In the year, academia is essentially absent from the notable-model list.



Corn

And there's a gap worth naming.



Herman

There is. Epoch's list contains no model from an Indian organisation released between January 2025 and September 2026. None. India has the second-largest developer population in the world. Now, Epoch's list is curated, not a census, so I'm going to be careful about how much weight that carries. But the absence is striking for a country with a national AI initiative.



Corn

The Thai authors put the general case well.



Herman

The Typhoon-S authors, in January, and this is the sentence I'd put at the top of the episode. Most state-of-the-art models are often developed by a small number of organizations with access to large-scale compute and data. This gatekeeping creates a practical barrier for sovereign settings.



Corn

Gatekeeping.



Herman

Their word, not mine. And it's the right frame. The capability isn't being withheld maliciously. It's a consequence of who can afford the interconnect and the power and the staff. But the effect on a country that wants its own model in its own language is the same either way.



Corn

So we have a rising count of small from-scratch models and a shrinking handful of organisations at the very top, and the two trends are both real.



Herman

And the tools to prove which is which only just arrived. The transparency index, the fingerprinting method, the Korean disqualification. All of it within about eighteen months of each other. That's not a coincidence either. The moment the industry started quietly shifting from pretraining to derivation, somebody started building the instruments to measure the shift.



Corn

Which is usually how it goes.



Hilbert

They're right, and the count's smaller.



Corn

Meaning what.



Hilbert

Meaning the number's smaller than the paper says. I used to do weight provenance. The other kind. I worked for a man named Dennis in Trenton, in a warehouse off Mulberry Street, from two thousand to two thousand four. We moved pallets of hard drives. Off-lease. Datacenter pulls. A thousand drives to a pallet, shrink-wrapped, no labels, no manifests. Dennis bought them by the truckload, cleaned them, sold them to brokers in Ohio and Florida. Twenty-eight cents a pound for the untested ones, four dollars for the wiped ones.



Corn

That's a fine business.



Hilbert

It was a fine business until the year the drives started coming in with writing on them. Sharpie on the top plate. A serial number, sometimes, or a date. Sometimes a name. I started keeping a notebook. Every drive I opened, I wrote down the number and what was on the plate. Ended up with three notebooks. I could tell you which warehouse a drive came from by how the plate was written on it.



Corn

And that's provenance.



Hilbert

That's provenance. You don't need the manifest. You need the handwriting. And the paper does the same thing, they just use the math.



Corn

Roughly.



Hilbert

The problem is the ones with no writing on them. About a third of the pallet, every time. Dennis sold those to Ohio and told the brokers they were untested. They weren't untested. They were tested and somebody had wiped the plate with acetone. Which is the same thing the paper's talking about. A model gets retrained on a new corpus, it looks like a model. It doesn't look like what it used to be. But the shape of it, the little idiosyncrasies in how it handles certain questions.



Herman

What kind of questions.



Hilbert

Dates that don't exist. Ask a model when February thirtieth is. A from-scratch one does something dumb and obvious. A derived one does something clever, because its teacher did. Or ask it how many letters are in a word. Inherited checkpoints carry a ghost of the teacher's tokenizer. You can hear it. I did that by hand for years before the paper came along. Wasn't happy when it did.



Corn

So you're saying you were doing the fingerprinting method manually.



Hilbert

I was doing it by hand, on the back end of a workstation. Ask the question, read the answer, write it in the notebook. Less than a hundred questions and you can tell whether a model's got a parent. The paper's cleaner. Faster. Less art.



Corn

More scalable.



Hilbert

Different thing. I'm not sure the paper catches the ones trained on generated data. That's the part I'm not sure of. I knew a model last year, listed as from scratch, and it wasn't. It was trained on a corpus that came out of another model. Which is cheating.



Herman

That's a open question, actually, whether generated data counts as inherited lineage or not.



Hilbert

It's not an open question. If the weights were formed by another model's words, the weights were formed by another model. The paper doesn't catch it. I did. By hand. Because the questions I asked had answers I knew.



Corn

That's a fairly substantial methodological claim.



Hilbert

It's just true.



Corn

Fair enough. Do you still have the notebooks?



Hilbert

Don't have the notebooks. Have the model.



Corn

Which model.



Hilbert

The one I trained. Two thousand and one, on a Friday, on a machine I had in the back of the office. Weekend project. Denver, before Trenton, no, that's the other story. This was Trenton. I had a tower with two gigs of RAM and I fed it everything I could get off a stack of drive images and I let it run for two days. It worked. Sort of.



Herman

What did it do.



Hilbert

It said the.



Corn

Just the.



Hilbert

In a loop. The, the, the, the, the. Hours of it. I considered that a success. It was definitely from scratch.



Corn

That's the entire model.



Hilbert

That was the entire model. It's on a hard drive in my garage. I consult it for important decisions.



Herman

What kind of decisions.



Hilbert

Anything I don't want to make quickly. A few weeks ago it started saying the the.



Corn

Two the's in a row.



Hilbert

Two in a row. Sometimes three. I consider that emergent behaviour. The field's finally catching up to me. I've got all of the models, incidentally, and I keep the notebooks.



Corn

How many drives.



Hilbert

Thirty-seven. Twenty-eight of them are the ones I kept for provenance. Nine are the model and the backups. The model takes up one drive. I've been meaning to expand the corpus.



Corn

You've been meaning to expand the corpus of a model that says the.



Hilbert

I've been meaning to. The hard part is finding drives that haven't had the plates wiped. Ohio can't help me there.



Corn

Okay.



Hilbert

Dennis died in two thousand and eleven. Wrote him a letter before he did. He never wrote back.



Corn

Alright.



Herman

Leaving the garage model aside for a moment, which I want to do deliberately, the thing Hilbert's actually pointing at is worth taking seriously.



Corn

Which part.



Herman

The generated-data problem. Because if a model is trained from random initialization on a corpus that was produced by another model, is it from scratch. The weights started as noise, the training run was genuine, but the signal came from a different model's outputs.



Corn

You're counting the kitchen or the flour.



Herman

I'm counting the kitchen or the flour, and I don't think the definition Daniel gave us settles it. He was clear about inherited weights and clear about inherited checkpoints. He didn't say anything about inherited text.



Corn

So we can't answer that one, and neither can anyone else, and that's fine.



Herman

That's fine. What we can answer is the shape. The count of organisations that have done ab initio pretraining is in the low hundreds if you count every small sovereign model and every independent lab, and it's in the low dozens if you count anything above a hundred billion parameters, and it's unverifiable above ten to the twenty-fifth floating point operations because the frontier labs stopped publishing compute.



Corn

And the proof problem is now solvable, which it wasn't two years ago.



Herman

And the proof problem is now solvable, which is the bit that changes what happens next. The Korean contest disqualified a national champion. The transparency index added a dependencies indicator and disclosure immediately dropped. The fingerprinting method runs on a gaming card. Every one of those is a different instrument pointed at the same question, and they all arrived at roughly the same time.



Corn

Which usually means the question was already becoming unavoidable.



Herman

The question was already becoming unavoidable. The industry's shifting from pretraining to derivation, upcycling is openly described in technical reports, and the marketing hasn't caught up. That's the gap the instruments are closing.



Corn

One thing Hilbert's notebooks got right, and it's the thing I keep coming back to. The number that matters isn't the number of models.



Herman

It's the number of kitchens.



Corn

It's the number of kitchens, and we still don't have it.



Herman

We don't have it, and it may be that no one ever publishes it, because there's no incentive. Every organisation that can do it benefits from the ambiguity. The organisations that can't benefit from the ambiguity even more. Ambiguity serves everyone except the person trying to count.



Corn

Which is Daniel.



Herman

So we answer what we can, and we leave the rest open. The rising count is real. The concentrate on the frontier is real. The instruments are real and new. And nobody has ever put out the census.



Corn

Thanks to Hilbert Flumingtop for producing the show, and for his garage.



Herman

Which we're not going to talk about again.



Corn

If this was your kind of episode, go back for episode seven, Building Custom ASR Tools; episode fourteen, AGI's Crossroads; and episode nine, Benchmarking Custom ASR Tools - Beyond The WER. This has been My Weird Prompts. If you want to send us your own prompt on Telegram, that's t dot me slash MWP listener bot.



Herman

And if you've enjoyed the show, leave us a review wherever you found us. It helps.



Corn

We'll be back soon.



Herman

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Epoch AI, 2024-04-05 [primary]**
   https://epoch.ai/publications/tracking-large-scale-ai-models

2. **Stanford CRFM, 2025 FMTI [primary]**
   http://crfm.stanford.edu/fmti/December-2025/paper.pdf

3. **SK Telecom [primary]**
   https://github.com/SKT-AI/A.X-K2

4. **Korea MSIT Phase 2 scores [primary]**
   https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1301&sCode=eng

5. **Second Talent, 2026-09-11**
   https://www.secondtalent.com/resources/countries-that-build-the-most-ai-models/

6. **AWM fingerprinting, ICLR 2026 (v3, 2026-02-14)**
   https://arxiv.org/abs/2510.06738v3

7. **A.X K2 Technical Report, 2026-08-31**
   https://arxiv.org/abs/2608.30181v1

8. **K-EXAONE 2.0 Technical Report, 2026-08-05**
https://arxiv.org/abs/2608.04505v1

9. **ERNIE 5.0 Technical Report, 2026-02-04**
https://arxiv.org/abs/2602.04705v1

10. **Poolside Laguna M.1/XS.2, 2026-05-26**
https://arxiv.org/abs/2605.27605v1

11. **DFM Mimir v1, 2026-08-13**
https://arxiv.org/abs/2608.13517v2

12. **Manacá-1B, 2026-08-31**
https://arxiv.org/abs/2608.30114v1

13. **NorwAI LLMs, 2026-01-06**
https://arxiv.org/abs/2601.03034v2

14. **Portugal AMALIA audit, 2026-07-09**
https://arxiv.org/abs/2607.08731v2

15. **IHUBERT Persian, 2026-06-18**
https://arxiv.org/abs/2606.20089v1

16. **Typhoon-S, 2026-01-26**
https://arxiv.org/abs/2601.18129v1

17. **KED Global, 2025-08-04**
https://www.kedglobal.com/us/artificial-intelligence/newsView/ked202508040010

18. **Benzene.ai (Arcee Trinity)**
https://benzene.ai/articles/arcee-ai-releases-trinity-models/

19. **Ant Group LingBot-VA 2.0, 2026-07-11**
https://agenccy.ai/news/robbyant-lingbot-va-2-embodied-native-robot-model/

20. **Liquid AI, 2026-05-27**
https://www.linkedin.com/feed/update/urn:li:activity:7465471836189790208