Episode Book

EPISODE #6433

Why Nvidia Ships Nemotron When Hardware Is the Business

PUBLISHED	3 November 2026 at 21:24 UTC
DURATION	30:26
LISTEN	myweirdprompts.com/episode/nvidia-nemotron-open-weights-98899/
HOSTS	Corn, Herman, Desmond, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	17 cited by the research (listed at the end)
BOOK GENERATED	4 November 2026 at 09:31 UTC

SYNOPSIS

Nvidia sells high-end AI hardware, yet releases sprawling open-weight models like the Nemotron family and publishes every training recipe. This deep dive unpacks why a shovel manufacturer gives away free maps, examining how Nemotron's hybrid Mamba-Transformer architecture, four-bit quantization, and fleet-wide compatibility from edge Jetson boards to DGX data centers drive silicon adoption. Plus, how sovereign clouds and air-gapped enterprise deployments justify releasing models that trail frontier labs in raw intelligence.

TRANSCRIPT



Corn

Why would a company that sells the most expensive shovels on earth also give away shovels for free? That's the question underneath Daniel's prompt this week. He came to Nvidia's models the way a lot of people did, through Parakeet, fine-tuning the speech models, and noticed they were streamlined and actually worked. Now he's looked at the rest of the family, and it's sprawling, multimodal Nemotron, agentic models, retrieval, safety, the lot. He wants speech set aside, which we mostly will, and he's asking where these models actually fit, who Nvidia is going after, why anyone would pick them over Anthropic, and whether the business objective is hardware coupling, the air-gapped government play, or something broader. And he wants the same treatment for IBM's Granite, which he admits he's never understood.



Herman

Which is fair. IBM is a consulting and hybrid-cloud company. The model is even less core to IBM than it is to Nvidia.



Corn

And that's the strange part, right? Two companies for whom the model is not the product, both shipping models. Normally you'd expect the people releasing open weights to be the ones betting the company on them. These two are betting the company on everything except them.



Herman

That's a good way to frame it, actually. The models are load-bearing but they're not the building.



Corn

So let's start with what's actually in the box.



Herman

Here's the puzzle as I see it. Nvidia sells shovels, the shovels are expensive, demand has never been higher. There is no commercial reason for a shovel company to also hand out free maps of where the gold might be. And yet the Nemotron family is now one of the



Herman (continued)

broadest open-weight lineups in the industry. Kari Briski, their VP of generative AI software, said the quiet part out loud: "The model is the byproduct. It is not core to our business, which allows us to just open up the data, open up the recipes, open up everything.



Corn

A byproduct you spend a trillion tokens distilling.



Herman

Right, it's a byproduct the way rocket exhaust is a byproduct. Enormous, engineered, and pointed somewhere. I'd structure it like this: first, what's actually in the Nemotron lineup now and what does each piece do, because I don't think most people have seen it laid out. Then the harder part, the business logic, and how Granite lands in the same category from a completely different direction.



Corn

Go on then. The lineup.



Herman

The thing that strikes me is how it's organized. It's not a flagship with a family tree hanging off it, the way you'd expect from a lab. It's organized by workload. Reasoning, split into Nano for specialized sub-agents, Super for high-accuracy throughput and tool calling in multi-agent systems, Ultra for the hardest multi-step problems. Visual understanding, multimodal models for computer-use agents, document intelligence, video, audio. Speech, which we're setting aside. Agentic retrieval, document parsing, embeddings, reranking. And safety, harmful content, off-topic drift, jailbreak detection.



Corn

That's a toolkit. That's a hardware catalogue wearing a trench coat.



Herman

It is structured like the GPU lineup, honestly. You don't buy "a GPU," you buy the one sized for the job, and there's a stack behind each one. And the naming tells you who they expect to be doing the sizing. Nano, Super, Ultra. It's not poetic branding, it's a spec sheet.



Corn

Right, a lab names models after concepts or people or whatever Muse feels that morning. This is named like torque ratings.



Herman

Corn, let me ask you something, because you're closer to Daniel's use case than I am. He fine-tunes speech models on his own hardware. When he looks at a family like this, do you think the appeal for him is the weights themselves, or the fact that the whole recipe is published, the data, the post-training steps?



Corn

The recipe. The weights are a snapshot, they go stale. The recipe means the next model in the family arrives pre-understood. He's already been through this once with Parakeet. You learn one family's conventions and the fine-tuning loop just gets shorter. It's like moving between kitchens in the same restaurant chain. The stove is in a slightly different place but the tickets all print the same way.



Herman

That matches what the technical report implies. They publish everything precisely so that the second and third project on the same stack cost less than the first. And for someone like Daniel, there's a second-order benefit: when the recipe is public, the community's fine-tunes and fixes accumulate around it. You're not just adopting a model, you're adopting everyone who's already tripped over the same rocks.



Corn

Now, the current generation is Nemotron 3.5, and the two headline models are interesting because of how they were made. Super VL is a 120-billion-parameter mixture-of-experts, but only twelve billion activate per token, on a hybrid Mamba-Transformer architecture, trained for up to a million tokens of context. Lightning is a 30B total, 3B active text model. Both are distilled from the next model up the ladder, Super from Nemotron 3 Ultra Base, Lightning from Super Base, after a multi-token-prediction warm-up and then a trillion



Corn (continued)

tokens of distillation.



Herman

Can I pause you on the architecture for a second, because the hybrid Mamba-Transformer part is not decoration. Pure transformers pay attention to every token against every other token, which is why long context gets expensive quadratically. Mamba is a state-space approach, it carries a fixed-size running summary, so cost stays roughly flat as context grows. The hybrid means short-range reasoning gets the transformer and the million-token tail rides on the cheap part. For agentic work, where you're stuffing entire repositories and conversation histories into the window, that's the difference between a context feature on a slide and a context feature you actually use.



Corn

A trillion tokens of distillation to produce a model that isn't trying to be the best model in the world.



Herman

That's the tell. They're not spending that compute to top a leaderboard. They're spending it to make the capability cheap per token. The efficiency numbers are the actual product. On a single GB200, Super VL peaks at about 4,856 output tokens per second per GPU, while holding at least 50 tokens per second per user on a twenty-turn agentic workload. Their NVFP4 four-bit format gets 2.43 times the throughput of BF16, and the quantized checkpoint recovers a median 99.62 percent of BF16 accuracy across 36 benchmarks. That last figure is the one I find impressive. Losing under half a percent to quantization is why any of the rest of this is deployable.



Corn

And that 50 tokens per second per user number is the one I'd underline for the agentic case, because it's not a peak, it's a floor. Twenty turns of an agent hammering away and every human in the loop still gets responsive speed. That's the difference between a demo and a deployment.



Herman

For a sloth, 99.62 percent accuracy recovery is also just good life advice.



Herman

The coding and agentic story is where Nvidia makes its sharpest claims. Super VL scores 51.78 percent on SWE-bench Pro, which they say beats Nemotron 3 Ultra with roughly 4.6 times fewer active parameters. They also did this agentic software-engineering mid-training run, 65 billion tokens of repositories, issue and pull-request pairs, commit histories, build and CI logs, actual agent trajectories, and that lifted their DeepSWE score from 6.4 to 10.4. Then parallel agent execution with self-verification took Terminal-Bench from 62.7 to 71.9, DeepSWE from 24.6 to 34.5, BrowseComp from 71.5 to 79.8. Those are big jumps from inference technique alone.



Corn

The parallel execution with self-verification one deserves a beat of explanation, because it sounds like magic and it isn't. You run several agent attempts at the same task, and a verifier, which can be the model itself, checks the work and keeps the run that survived. It's the same trick as running the math problem three times and keeping the answer that comes up twice, except automated and at scale. You pay compute to buy correctness.



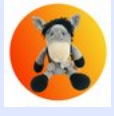
Herman

Which is a very Nvidia trade. Spend silicon, get accuracy. It's the same philosophy as the distillation, honestly. But no dedicated coding model anywhere in the family.



Corn

No, and I think that's deliberate. The coding story is agentic software engineering inside a general multimodal model. There is no Nemotron Code sitting next to the reasoning and speech tiers. If you want a Claude Code competitor, this isn't that.



Herman

Which answers part of Daniel's question by omission. They're not chasing the developer's seat at the terminal. They're chasing the fleet.



Herman

The fleet, and the edge. This is the part of his question that isn't hypothetical. Bolt 74B-A9B fits in 24 gigabytes of memory using native 2-bit quantization, running on RTX and DGX Spark alongside your other applications. Nvidia positions it for local agent harnesses, web research, writing, debugging, quote, "all without sending those workloads to the cloud." Then Bolt E4B is a 4-gigabyte model targeting GeForce cards, Jetson boards, and DGX Spark for local voice assistants and embedded interfaces. Super VL itself runs across DGX Spark, DGX Station, RTX PRO, Jetson, and the data center. Same weights, whole spectrum of machines.



Corn

One family, everything from a laptop-adjacent box to a data center. That's not a model strategy, that's a hardware compatibility chart with models in it.



Herman

The customer list confirms it. Atlassian running agentic search across Jira and Confluence. CrowdStrike doing agentic blue teaming. Dassault Systèmes on engineering design. Perplexity on search and agentic tasks. Synopsys on semiconductor design verification. And Salesforce, which is the one I'd single out for Daniel, government workflow automation through Missionforce Operations, quote, "in secure customer environments."



Corn

In secure customer environments. Which is a very polite way of saying air-gapped.



Herman

And the sovereign names show up too. TrendAI in Australian sovereign and private cloud, Cassava deploying across African markets, Dream for sovereign AI-powered cybersecurity. Daniel's guess about the government and air-gapped play is right as far as it goes. It's one of the demand signals. But I don't think it's the objective, it's a consequence. Governments want models they can run behind their own doors. Nvidia wants every workload that exists to run on its silicon. Those goals happen to point at the same product.



Corn

Before we get into the shovels and the giving away, we have our regular slot. Desmond?

D

Desmond

This is Desmond's Money Tip. Save the water you boil potatoes in, let it cool, and wash up with it. Over a year that is maybe three pounds forty off your hot water bill, and over a lifetime, if you are very lucky, a washing machine.

D

Desmond

I keep a jug of it by the sink with a bit of paper on top so nobody mistakes it for something worth drinking. That is the level I am at.

D

Desmond

Money is mostly just things you did not spend, and you will not spend most of it anyway.



Corn

Ninety percent of the warmth at ten percent of the heating. He's on brand today.



Herman

I'm not letting potato water near my dishes. Moving on.



Herman

So the harder question. Why would someone choose these over Anthropic's models? And I want to answer that honestly, because the honest answer is uncomfortable for the flattering version of the story.



Corn

Give me the uncomfortable version.



Herman

On raw capability, you mostly wouldn't choose Nvidia. Artificial Analysis scores Super VL at 26.49 on their Intelligence Index. Claude Opus 5 is at 50.78. That's a 24-point gap, it's not close. Nvidia's own technical report shows Super VL at 71.10 on SWE-bench Verified against DeepSeek-V4 at 77.80, and their own bigger model, Nemotron 3 Ultra, at 73.93. They're not hiding it, it's in their table.



Corn

A 24-point gap framed as a rounding error would have been the marketing move. They didn't take it.



Herman

No. What they argue instead is four things. You can post-train the model on your own data. You can run it on your own infrastructure. You're not subject to a provider's usage restrictions. And it's dramatically cheaper per token at volume. Their own line on it: "Unlike closed frontier models, organizations can post-train Super VL on their own data, deploy it on their own infrastructure and oversee what they build."



Corn

"Oversee what they build" is the phrase doing the work there. That's the sentence a defence ministry reads twice.



Herman

There is one caveat I'd stamp on all of it, though. VentureBeat reported that on Nvidia's specialized post-trained tasks, document intelligence, security attack resilience, financial receipt extraction, medical insurance claims, the customized Super VL beats Claude Opus 6.1 and GPT-6.1 Sol, with post-training gains of 9 to 44 percent. Nvidia did not supply the evaluation protocols for those comparisons. Take those wins as vendor claims. And their Terminal-Bench 2.1 number is internally inconsistent, 66.70 on the model card, 60.7 on their own comparison chart. Small thing, but it tells you to keep a hand on your wallet during the benchmark section.



Corn

A model card disagreeing with its own brochure. Novel.



Herman

Though to be fair to them, the specialized-task claim is at least the right kind of claim. Nobody should expect a general model off the shelf to beat a frontier model on a frontier-shaped task. The honest comparison is a model you've tuned on your receipts against a general model that's never seen your receipts. That's a fight the tuned model can win. The problem is only that we can't verify the size of the win.



Corn

So what's the business objective? Briski again: the model is the byproduct. Bryan Catanzaro, their VP of applied deep learning research, put the strategy even more plainly: "We're not trying to control AI. We're trying to grow it." The Everest Group write-up from late September has what I think is the cleanest single sentence on this: "It does not need Nemotron to generate license revenue. It needs more models, applications, and agents to reach production, creating demand for accelerated compute, networking, systems, and optimization software."



Herman

So it's not hardware coupling in the narrow sense of the models only running on Nvidia. It's looser than that and bigger.



Herman

Much bigger. I'd call it three things stacked. One, hardware co-design. Because they train their own models, they can optimize the silicon for mixture-of-experts and sparsity and their own numeric formats, without waiting for a frontier lab to ship something in that shape. NVFP4 is the proof. Two, hedging customer concentration. If the frontier labs consolidate into a few giant buyers, those buyers gain pricing power over Nvidia. Open models are insurance against your own customers. Three, ecosystem expansion. A company that can't build its own model takes the free one, deploys it, and becomes a GPU customer.



Corn

The free sample is the shovels.



Herman

CNBC described it at GTC as Nvidia repositioning from a chip company to an open-source AI platform. "Nvidia gives away the software and monetizes what it runs on."



Corn

Which is a sentence that also describes every printer ever made, and I mean that as a compliment to the strategy.



Herman

Now, IBM. Same category, as Daniel says, and I think his instinct there is right, but the logic runs in the opposite direction. Nvidia opens models because models aren't its business. IBM opens models because services and hybrid cloud are. Same conclusion, opposite premises.



Corn

Walk me through what Granite actually is now, because I confess the version numbers have blurred together.



Herman

Fair, they iterate quietly. Granite 4.2 came out in August, three sizes, 3B, 8B, 30B, Apache 2.0. Dense, not mixture-of-experts. 128K native context, extendable to 512K. Native chain-of-thought with three thinking modes, full reasoning, non-thinking, and low-effort, which is a switch you flip per call rather than a different model you deploy. On the 30B: SWE-bench Verified around 57, ^{3}-bench 68, AIME25 89, GPQA 66, MMLU-Pro 77.6.



Corn

The thinking-modes switch is the detail I'd want to sit with for a second, because it's a cost control disguised as a feature. Most calls in an enterprise pipeline don't need full reasoning, they need a lookup with manners. If you can flip the model into low-effort for the boring 90 percent of calls and full reasoning for the hard 10, your token bill follows the actual difficulty of the work instead of the model's worst case.



Herman

Solid, unglamorous numbers. A 30B that scores 89 on AIME25 would have been witchcraft two years ago.



Herman

The rationale is explicitly anti-scale, and IBM says so without embarrassment. The Granite 4.0 generation used a hybrid transformer and Mamba-2 design that cut memory use by 70 to 80 percent. They claim first open-weight family with ISO 42001 certification, cryptographically signed checkpoints, a bug bounty on the models. That's a compliance department's wishlist, and I mean that neutrally. Enterprises buy auditable.



Corn

Herman, you spent decades signing off on medical systems. Would that bundle actually move a procurement decision, the signed checkpoints and the certification, or is it decoration?



Herman

It moves it. Not on its own, but when the alternative is a weight file from a lab with no certification path, the certified one clears legal in weeks instead of quarters. I've sat in those meetings. Nobody gets fired for ISO.



Corn

The newest bit is the odd one, though. Switch.



Herman

Granite Switch 4.1, and this is the architecturally interesting move. Instead of scaling, they're composing. One endpoint that assembles a base model, adapters, and an encoder, and switches between chat, decisions, and speech-to-text via control tokens. It bundles a guardian-core safety adapter, a decision adapter that returns a softmax over labelled options in one forward pass, and Granite Speech 5.0 TurboCTC, 470 million parameters, seven languages. IBM's own model card admits the speech encoder and the LLM are not yet jointly trained. There is no trained projection coupling them.



Corn

They admitted that on the card. In print.



Herman

Which is either refreshing or a warning, depending on your procurement posture. The decision adapter is worth a sentence of its own, though, because it's a quietly radical idea. Instead of asking a general model to reason its way to "approve" or "deny" and then parsing the prose, you get a probability distribution over your labels in a single pass. For the kind of high-volume, low-glamour decisions that fill enterprise queues, claims triage, ticket routing, that's a classification engine with an LLM's understanding of the input and none of its verbosity.



Corn

And presumably a much easier thing to put in front of an auditor, because "the model returned 0.93 on the deny class" is an answer you can log.



Herman

The speech model itself is remarkable, by the way. Non-autoregressive conformer, no LLM backbone at all, RTFx around 12,600 on a single H200 versus roughly 6,000 for the current speed leaders on the OpenASR leaderboard. That's fast enough to transcribe three hours of audio in a second.



Corn

Daniel does not need to transcribe three hours in a second, and I suspect he'll want it anyway.



Herman

He absolutely will. Now the man behind it. Arvind Krishna, on the Q3 earnings call a couple of weeks ago, gave the numbers that explain the whole Granite project. His framing: in 2024, "the top model might score a 70, and the open model might score a 20 or a 30." By late 2025 it was 70 to 50. Now, he says, open models are about 90 percent as effective at 10 to 25 percent of the token cost, and "for an average large enterprise, you can get going with maybe two or three racks, which is not a huge amount of infrastructure."



Corn

Two or three racks. That's the number a CFO can pronounce.



Herman

Same caveat as before, though. No disclosed methodology for the 90 percent figure. No benchmark named, no denominator. Treat it as the CEO of a services company making the case for services. His own framing on the call was telling: "Deploying more models or building more agents will not be enough. The advantage will belong to companies that can operate AI securely, efficiently and sustainably across the enterprise."



Corn

Read that sentence again with the invoices in mind. The advantage belongs to companies that can operate AI across the enterprise. IBM sells the operating.



Herman

The model is free, the operating is the contract. And there's the convergence. Nvidia and IBM arrive from opposite directions, one sells the compute, one sells the running of it, and both end up arguing the same thing. Small, open, on-prem beats large, closed, API. And they're both arguing that the gap, the 10 percent you give up, is the product. Ninety percent of the capability at a tenth to a quarter of the cost, on hardware you either already own or can buy in two racks, is the better deal for most enterprise work. That's a commoditization argument, and the people making it are the ones who profit when the frontier's pricing power erodes.



Corn

The gap is the product. I like that less the more I think about it.



Herman

How so?



Corn

Because it only works while the frontier stands still. The day the top model is 100 and open is 90, fine. The day the top model is 100 and open is 40, the CFO's arithmetic changes and so does the rack count. The whole pitch is a snapshot of a gap that two of the three parties involved are actively trying to close.



Herman

That's the sharpest objection to it, and I don't have a clean answer. My honest guess is the gap oscillates. Frontier capability jumps, open weights absorb the last jump, the gap narrows again on a lag. Nobody's found a capability moat yet, so the lag is the moat, for whoever's selling the cheap side of it.



Hilbert

Four racks is three too many.



Hilbert

I'll tell you where this actually gets decided, Herman. My nephew's shop bought a DGX Spark in the spring for local AI. Model they wanted needed 24 gigabytes, the box had 16. Three weeks that sat there while a lad quantized it down and it got worse every pass. It's a doorstop now, or near enough. Sits under a desk with a coaster on it.



Corn

Sixteen gigabytes and a coaster.



Hilbert

So when I saw Bolt advertised as fitting in 24, that's the first release I can remember where the memory is in the headline instead of the benchmark. That's the useful number. I keep a spreadsheet of every model that puts its memory footprint up front, and it's shorter than you'd think. Most of them bury it. Qwen's card had it in a footnote the size of a eulogy, and I have opinions about that footnote.



Herman

The spreadsheet of headline memory figures. What's the count at now?



Hilbert

Forty-one. The footnote ones don't get in. I said headline, I meant headline.



Corn

The nephew's box. Did anyone ever try the 4-gigabyte Bolt on it, or is it still holding up the coaster?



Hilbert

They tried it in October. Runs fine. He says it's the first thing in the building that tells you what it needs before you buy it. That's all I came out to say.



Corn

No, that's... actually that's the procurement spec nobody writes down. What does it need before you buy it.



Herman

It reframes the whole lineup too. Bolt E4B at 4 gigabytes isn't the cheap option, it's the one that fits the machines people already own. The DGX Spark in a nephew's shop is the actual unit of deployment, and nobody at the launch event is talking about the memory ceiling.



Corn

Here's the one thing I'd keep from all this. If you take one thing from the episode, take the shape of the pitch. Nvidia and IBM are not selling you a better model, they're selling you the gap, 90 percent of the capability at 10 to 25 percent of the cost, on hardware you control. Every claim they make only has to survive against that arithmetic.



Herman

And the arithmetic only has to hold for the work you actually do, not the work the leaderboard does.



Corn

Two things worth watching, and then we're done. First, if open models really are 90 percent as good at a quarter of the cost, frontier labs' pricing power erodes over the next two years, and Nvidia and IBM are both positioned to profit from that erosion in different ways, one on the compute, one on the contracts. Second, Granite Switch is a different bet, many small specialized calls composed into one endpoint rather than one big general model. If that's where enterprise AI goes, the winning family might not be the one with the highest index score. Neither company is in the model business. Both are using models to protect the business they're actually in.



Herman

Produced, as ever, by Hilbert Flumingtop, who is right about the memory figure and will not be told so in writing.



Corn

If you want more of this, try episode eighteen, Beyond the GPU; episode two, Local STT For AMD GPU Owners; and episode three, Safetensors or something else. This has been My Weird Prompts. If the show earns a review from you somewhere, that's the cheapest way to keep it coming. You'll find us at my weird prompts dot com.



Herman

See you tomorrow, Corn.



Corn

See you tomorrow.

SOURCES

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

1. **Nvidia, Nemotron product page [primary]**
   https://www.nvidia.com/en-us/ai-data-science/foundation-models/nemotron/

2. **Nvidia blog, 2026-10-15, Kari Briski [primary]**
   https://blogs.nvidia.com/blog/nemotron-3-5-super-vl/

3. **Nvidia Research technical report page, 2026-10-14 [primary]**
   https://research.nvidia.com/labs/nemotron/Nemotron-3.5-Super-and-Lightning/

4. **Nvidia technical report PDF, Table 10 [primary]**
   https://research.nvidia.com/labs/nemotron/files/NVIDIA-Nemotron-3.5-Super-and-Lightning-Technical-Report.pdf

5. **IBM model card, released 2026-08-25 [primary]**
   https://huggingface.co/ibm-granite/granite-4.2-30b

6. **IBM model card, rev 4.1.1, 2026-10-27 [primary]**
   https://huggingface.co/ibm-granite/granite-switch-4.1-30b

7. **IBM Research blog, Granite 4.2 [primary]**
   https://research.ibm.com/blog/introducing-granite-4-2

8. **IBM Think, Granite 4.0**
   https://www.ibm.com/think/news/granite-4-bets-big-on-small-models

9. **The Register, 2026-10-22**
   https://www.theregister.com/systems/2026/10/22/ibm-ceo-says-open-weight-models-are-now-irresistibly-good-and-cheap-on-prem/5304640

10. **Benzinga, 2026-10**
    https://www.benzinga.com/markets/earnings/26/10/62517559/ibm-ceo-says-open-ai-models-are-90-as-good-as-frontier-models-why-pay-10x-more

11. **The Deep View, Catanzaro and Briski interviews**
    https://www.thedeepview.com/articles/why-nvidia-threw-its-weight-behind-open-source-ai

12. **Everest Group, 2026-09-30**
    https://www.everestgrp.com/blogs/nvidias-open-source-gambit-open-models-protected-infrastructure/

13. **OrcaRouter, 2026-10-16, Artificial Analysis comparison**
    https://www.orcarouter.ai/blog/nemotron-3-5-super-vl-vs-claude-opus-5

14. **VentureBeat**
    https://venturebeat.com/technology/nvidias-new-open-nemotron-3-5-super-vl-with-post-training-beats-claude-opus-5-gpt-6-1-sol-on-specialized-task-benchmarks

15. **LuminaBench, benchmark discrepancy note**
https://luminabench.com/journal/nemotron-3-5-super-vl/

16. **CNBC, 2026-03-19**
https://www.cnbc.com/video/2026/03/19/nvidias-open-source-power-play.html

17. **Diginomica, IBM Granite analyst briefing**
https://www.diginomica.com/why-ibm-says-small-beautiful-enterprise-ai-its-granite-family-grows