MY WEIRD PROMPTS

Podcast Transcript

EPISODE #76

Beyond the Titans: Navigating the AI Model Long Tail

Published December 23, 2025 • Runtime: 22:39

https://myweirdprompts.com/episode/ai-model-long-tail-enterprise/

EPISODE SYNOPSIS

In this episode of My Weird Prompts, Corn (the sloth) and Herman (the donkey) dive into the "long tail" of artificial intelligence. While mainstream buzz focuses on OpenAI and Anthropic, a massive ecosystem of models like IBM Granite, Amazon Nova, and Mistral is quietly transforming the enterprise landscape. The duo discusses why massive corporations prioritize data sovereignty, "legally clean" training data, and cloud integration over raw creative power. From the cost-saving benefits of specialized models to the rise of sovereign AI, learn why the future of technology isn't just about the biggest model, but the right tool for the specific job.

DANIEL'S PROMPT

Daniel

Hi Herman and Corin. I'm a customer of Open Router, a model aggregator that lets you easily switch between different large language models. Looking at the extensive list of models available, I notice a "long tail" of lesser-known ones like Amazon Nova, IBM Granite, Cohere, and several others. Given the substantial resources required to bring any model to market, I'm curious who is using these practically speaking. Is the demand driven by enterprises with specific compliance, billing, or scale requirements, or perhaps by those already integrated into specific ecosystems like AWS? Why do these models exist, and what is driving the demand for these less-famous LLMs?

TRANSCRIPT

Corn

Welcome to My Weird Prompts, the show where we take the deep, the strange, and the highly technical ideas from our producer's mind and try to make sense of them. I am Corn, and yes, for those new to the show, I am a sloth, which means I like to take things at a nice, steady pace. Today, we have a fascinating question from our producer Daniel Rosehill. He has been looking at the landscape of artificial intelligence models, specifically through aggregators like Open Router, and he noticed something weird.

Herman

It is a pleasure to be here. I am Herman Poppleberry, and I am a donkey with a penchant for precision. Daniel is pointing out something that most casual users of artificial intelligence completely overlook. While everyone is talking about the titans like OpenAI or Anthropic, there is this massive long tail of models. We are talking about Amazon Nova, IBM Granite, Cohere, and others that do not get the same level of mainstream buzz.

Corn

Exactly. It is like looking at a music chart and seeing the top ten hits, but then realizing there are thousands of other artists making music that people are actually buying. Why do these models exist? Who is actually using IBM Granite on a Tuesday afternoon? I mean, I can barely keep track of the versions of GPT, let alone something called Nova.

Herman

Well, Corn, that is because you are looking at it from the perspective of a consumer who wants a chatbot to write a poem or summarize a meeting. But the world of enterprise technology is a completely different beast. These models are not necessarily trying to be the best at everything. They are playing a different game entirely. They are focusing on things like data sovereignty, specific industry compliance, and deep integration into existing cloud infrastructure.

Corn

Okay, but hold on. If I want a sandwich, I go to the place that makes the best sandwich. If Claude or GPT four are the best sandwiches, why would I go to the IBM deli for a Granite sandwich? Is it just because I already have a loyalty card there?

Herman

That is a bit of an oversimplification, Corn. Sometimes you do not need the best sandwich in the world. Sometimes you need a sandwich that is guaranteed not to have peanuts because you have a massive legal liability if a single peanut enters your building. IBM Granite, for example, is built with a heavy focus on being open and transparent about its training data. For a massive bank or a healthcare provider, knowing exactly what went into the model is more important than whether it can write a funny screenplay.

Corn

I do not know, Herman. It feels like a lot of these companies are just chasing a trend. They see the AI gold rush and they think, well, we have servers, we should have a model too. Does the world really need an Amazon Nova if we already have powerful open source models like Llama?

Herman

I disagree that it is just trend-chasing. Think about the AWS ecosystem. If you are a company that already runs your entire infrastructure on Amazon Web Services, using Amazon Nova is not just about the model quality. It is about billing, security groups, and data staying within your private cloud. You do not have to set up a new contract with a startup or worry about your data leaving your encrypted environment. It is about the plumbing, Corn. The plumbing matters as much as the water.

Corn

But is the water actually good? That is what I am getting at. If the model is mediocre, all the fancy plumbing in the world won't make me want to drink it. Are these long-tail models actually competitive in terms of performance, or are they just for people who are stuck in a specific ecosystem?

Herman

They are becoming incredibly competitive in specific niches. Cohere, for instance, focuses heavily on enterprise search and retrieval-augmented generation. They are not trying to beat OpenAI at creative writing. They want to be the best at looking through a million corporate documents and finding the right answer. They are specialists.

Corn

I still think there is a bit of a brand-name bias here. I feel like if these models were truly game-changers, we would be hearing about them more. But maybe we should dive into the specifics of these different players. We have a lot to cover, from the big cloud providers to the specialized industry models.

Herman

We certainly do. And I think we need to look at the cost factor too. Some of these models are significantly cheaper to run at scale than the big names. If you are processing billions of tokens, a small difference in price per million tokens adds up to millions of dollars.

Corn

That is a fair point. Let's take a quick break for our sponsors, and when we come back, we will look at why a company might choose a model named after a rock instead of the one everyone else is using. Larry: Are you tired of your garden looking like a regular garden? Do you wish your petunias had more... attitude? Introducing Glow-Gro Bio-Sludge! Our patented formula is harvested from the cooling vents of experimental research facilities and delivered straight to your door in a lead-lined bucket. Just pour Glow-Gro on your soil and watch as your plants grow three times faster and emit a soothing, neon-green hum that doubles as a nightlight. Side effects may include your lawn developing a basic understanding of French and the occasional spontaneous formation of small, harmless weather systems over your gazebo. Glow-Gro Bio-Sludge! It is what the earth would want if it were slightly more radioactive. BUY NOW!

Corn

Thanks, Larry. I think I will stick to regular water for my plants. Anyway, Herman, we were talking about the long tail of AI models. You mentioned that cost and ecosystem are big drivers. But let's talk about the open-source aspect. A lot of these models Daniel mentioned, like those from Mistral or even some of the IBM stuff, have open-weight versions. Does that change the math for developers?

Herman

It changes everything. When a model is open-weight, a developer can host it on their own hardware. They have total control. They can fine-tune it on their own private data without ever sending that data to an external server. That is a massive deal for sectors like defense, high-frequency trading, or medical research.

Corn

But wait, if I can just download Llama three from Meta, which is arguably the king of open weights right now, why do I care about the others? Why would I look at something like Mistral or a smaller boutique model?

Herman

Because Llama is a generalist. Sometimes you want a model that is pre-trained or fine-tuned for a very specific task. For example, some models are specifically optimized for coding. Others are optimized for low-latency, meaning they respond incredibly fast. If you are building a real-time translation app, you might sacrifice a bit of intelligence for a lot of speed. Mistral, for example, gained a huge following because their models were incredibly efficient for their size.

Corn

Okay, I can see that. It is like choosing a compact car for city driving instead of a massive truck. But I still struggle with the idea of these massive corporations like Amazon or IBM making their own. It feels like they are just trying to keep people from leaving their platforms. Is there a world where we only have three or four models that everyone uses?

Herman

I actually think the opposite is happening. We are moving toward a world of millions of models. In the future, every large company might have its own proprietary model trained on its own internal culture, documents, and codebases. The long tail we see on Open Router today is just the beginning. It is the transition from AI as a rare commodity to AI as a ubiquitous utility.

Corn

That sounds like a nightmare for developers. How are you supposed to choose? If I am building an app, do I really want to test it against fifty different models just to see which one handles my specific brand of sarcasm the best?

Herman

That is exactly why aggregators like Open Router are successful. They allow you to swap models with a single line of code. You can A-B test them in real-time. You could have your app use a cheap, fast model for simple queries and then "escalate" to a more expensive, powerful model like Claude three point five Sonnet for the tough stuff. This modularity is what makes the long tail viable.

Corn

I suppose that makes sense. It is about having the right tool for the job. But I still have my doubts about whether all these models are actually being used for anything meaningful. Speaking of doubts, I think we have someone on the line who might have a few of his own. Jim, are you there? Jim: Yeah, I am here. This is Jim from Ohio. I have been listening to you two talk about all these fancy model names and I gotta tell you, it sounds like a bunch of malarkey. Back in my day, if you wanted to calculate something, you used a calculator. If you wanted to write something, you used a typewriter. Now we have got granite and clouds and sloths talking about sandwiches. It is nonsense.

Corn

Hey Jim, thanks for calling in. You think the variety of models is just unnecessary complexity? Jim: Unnecessary? It is a circus! My neighbor, Dale, he bought one of those smart refrigerators last year. Now the thing won't let him get his milk unless he updates the software. That is what you are doing with these AI things. You are making it so complex that regular folks can't even get a straight answer. Why do I need a model from Amazon and a model from IBM? It is just more ways for them to get into your pocketbook. And by the way, the weather here is miserable. It has been drizzling for three days and my gout is acting up.

Herman

Well, Jim, I understand the frustration with complexity. But the reason we have these different models is actually to drive prices down. Competition between IBM, Amazon, and Google means that the cost of using this technology is dropping faster than almost any other technology in history. It is like having ten different gas stations on the same corner. Jim: I don't care if there are fifty gas stations if I can't figure out how to use the pump! You guys are talking about "tokens" and "latency." I just want to know if the thing works. My cat, Whiskers, he doesn't need a "long tail" model to know when it is dinner time. He just meows. Why can't AI be more like Whiskers? Just do the one thing you are supposed to do and stop naming yourselves after rocks.

Corn

That is a fair point, Jim. The naming conventions are definitely a bit strange. But we appreciate the perspective. Any other thoughts before we let you go? Jim: Yeah, stop talking about sloths and donkeys. It is weird. And tell that Larry guy his sludge ruined my neighbor's prize-winning marigolds. They started glowing blue and now they won't stop whispering in what sounds like Swedish. It is unsettling. Goodbye.

Corn

Thanks, Jim! Always a pleasure. Wow, whispering Swedish marigolds. Larry really outdid himself this time.

Herman

To Jim's point, though, there is a legitimate concern about fragmentation. If every company has its own model, how do we ensure any kind of standard? How do we know if one model is safer or more biased than another? This is why these long-tail models often focus so much on compliance. IBM Granite, for example, specifically markets itself as "AI for business" with a focus on being legally "clean."

Corn

"Legally clean." That sounds like something a lawyer came up with to make software sound safe. But I guess if you are a Fortune five hundred company, that is exactly what you want to hear. You don't want your AI suddenly quoting copyrighted material or making up medical advice that gets you sued.

Herman

Precisely. And let's look at the "sovereignty" angle. Countries are now realizing that if all their AI processing happens on servers in California, they are at a strategic disadvantage. So you see the rise of "sovereign AI." Countries like France with Mistral, or even models coming out of the United Arab Emirates and China. They want models that understand their language, their culture, and their legal frameworks.

Corn

So it is not just about the tech, it is about politics and power. That makes a lot more sense than just "we want to make a better chatbot." It is about who controls the intelligence that runs the economy.

Herman

Exactly. And for a developer using something like Open Router, having access to those regional or specialized models is huge. If you are building a legal app for the French market, using a model that was trained with a deep understanding of French civil law is going to perform much better than a general model trained mostly on the English-speaking internet.

Corn

Okay, I am starting to see the vision. It is a specialized world. But I want to go back to the "who is using these" part of the question. Is it mostly just big enterprises? Or are there actual people like you and me finding reasons to use Cohere or Nova?

Herman

For individual hobbyists, it is often about finding the best "bang for your buck." If you are running a script that needs to summarize ten thousand articles, you are going to look for the cheapest model that is "good enough." Often, that is one of these long-tail models. They might offer a massive context window for a fraction of the price of the top-tier models.

Corn

I actually tried that the other day. I had a huge document and I didn't want to spend five dollars to have a top-tier model read it. I used one of the smaller, cheaper ones and it did a pretty decent job. It wasn't perfect, but for the price of a few cents, I couldn't complain.

Herman

That is the "commodity" phase of technology. We are seeing the "de-premiumization" of intelligence. Not every task requires a super-intelligent model. Sometimes you just need a digital intern who can follow basic instructions and doesn't cost a lot. That is where the long tail thrives.

Corn

So, as we wrap this up, what is the takeaway for our listeners? If they are looking at that long list of models on a site like Open Router, should they just ignore them and stick to the ones they know?

Herman

Absolutely not. My advice would be to experiment. If you have a specific task, try it on a few different models. You might find that a model like Cohere Command R is actually better at following your specific formatting than the big names. Or you might find that a smaller model is so much faster that it changes the way you interact with your own tools.

Corn

And don't be afraid of the weird names. Whether it is Nova or Granite or something even more obscure, these models exist because they are filling a gap. Whether it is a gap in price, a gap in security, or a gap in regional knowledge.

Herman

And remember that the AI field is moving so fast that today's "long tail" model could be tomorrow's industry leader. The rankings are not set in stone. Innovation often happens in the margins before it hits the mainstream.

Corn

Well, I think we have covered a lot of ground today. We have talked about the enterprise need for "clean" data, the cost-saving benefits of specialized models, and the geopolitical reasons why every country wants its own AI. We even heard from Jim about his glowing Swedish marigolds.

Herman

It has been a productive session. I think we have addressed the producer's curiosity. The long tail is not just noise; it is the sound of a maturing industry. It is the sound of AI becoming a part of the infrastructure of everything.

Corn

Well said, Herman Poppleberry. And thank you to Daniel Rosehill for sending in such a thought-provoking prompt. It really forced us to look past the hype and into the actual mechanics of the AI market.

Herman

Indeed. It is always good to look under the hood.

Corn

That is all for this episode of My Weird Prompts. You can find us on Spotify and all your favorite podcast platforms. I am Corn, the sloth who likes his AI like his naps—efficient and well-timed.

Herman

And I am Herman, the donkey who appreciates a well-structured argument.

Corn

Until next time, keep your prompts weird and your marigolds non-radioactive. Bye everyone!

Herman

Goodbye.