MY WEIRD PROMPTS

Podcast Transcript

EPISODE #107

The $5.5 Million Breakthrough: DeepSeek’s AI Disruption

Published December 26, 2025 • Runtime: 17:41

https://myweirdprompts.com/episode/deepseek-ai-efficiency-disruption/

EPISODE SYNOPSIS

In this episode of My Weird Prompts, Herman and Corn dive deep into the seismic shift occurring in the artificial intelligence landscape as Eastern models like DeepSeek and Z.ai challenge the status quo. While Western giants like OpenAI and Anthropic spend hundreds of millions on training, DeepSeek has managed to produce world-class performance for a mere $5.5 million. The duo explores the technical "wizardry" behind this efficiency, including Multi-Head Latent Attention (MLA) and FP8 mixed precision training, which allow these models to run on less expensive hardware without sacrificing power. They also tackle the strategic implications of open-sourcing these models under MIT licenses, the impact of hardware export bans on innovation, and how Western developers are increasingly turning to these cost-effective alternatives to build the next generation of apps. Is AI intelligence becoming a cheap commodity like electricity? Join Herman and Corn as they unpack the economic and technical forces turning the AI world upside down.

DANIEL'S PROMPT

Daniel

How can companies like Z.ai and DeepSeek offer their models at such a significant cost difference compared to American competitors like Anthropic and OpenAI? Is the price difference due to model size, training efficiencies, economic factors, or a strategic pricing move? Additionally, is there any data on the adoption of these Eastern models in the West?

TRANSCRIPT

Corn

Hey everyone, welcome back to My Weird Prompts! I am Corn, and as always, I am hanging out here in Jerusalem with my brother.

Herman

Herman Poppleberry, at your service. It is a beautiful day to talk about some seriously heavy-duty technology.

Corn

It really is. And we have got a great one today. Our housemate Daniel sent us a voice note about something that has been making waves in the tech world lately. He was asking about these artificial intelligence models coming out of the East, specifically companies like DeepSeek and Z dot A I.

Herman

Yeah, Daniel is always keeping his ear to the ground with this stuff. He noticed that these models are hitting the market with price tags that are just a fraction of what companies like OpenAI and Anthropic are charging. We are talking about a massive gap in cost.

Corn

It is wild. I mean, I have been seeing the headlines, but when you actually look at the numbers, it feels like there is a glitch in the matrix. How can one company charge thirty times less than another for something that performs almost as well, or sometimes even better?

Herman

That is the multi-million dollar question, Corn. Or in the case of DeepSeek, maybe the five point five million dollar question.

Corn

Wait, five point five million? That sounds like a lot of money to me, but in the world of training these massive models, that is actually peanuts, right?

Herman

Exactly. To put it in perspective, the industry standard for training a top-tier model like GPT-four was rumored to be well over one hundred million dollars. Some estimates for the next generation of Western models are heading into the billions. And then DeepSeek comes along and says, hey, we built DeepSeek-V-three for about five point five million dollars in training costs.

Corn

Okay, hold on. We need to back up. If I am a regular person just using these tools to write emails or code, why should I care about the training cost? And how on earth did they get it that low?

Herman

Well, you care because those savings get passed directly to you. When the training and inference costs are lower, the price per token, which is basically how these companies measure the text they generate, drops through the floor. It makes it possible for developers to build much more complex apps without going bankrupt.

Corn

That makes sense. So, is it just that they are using cheaper labor or something? Or is there actual wizardry happening under the hood?

Herman

It is a bit of both, but the technical wizardry is the real story here. DeepSeek-V-three uses some incredibly clever architectural tricks. One of the big ones is something called Multi-Head Latent Attention, or M L A.

Corn

Okay, Herman, you know the drill. Break that down for a sloth like me. What is Multi-Head Latent Attention when it is at home?

Herman

Think of the model's attention mechanism as its ability to look back at what has already been said to understand the context. In older models, that process takes up a huge amount of memory. It is like having a giant filing cabinet where every single word has its own massive folder. M L A is like a super-efficient compression system. It allows the model to keep all that context in a much smaller space, which means it can process information much faster and on less expensive hardware.

Corn

So it is like they found a way to pack the same amount of brainpower into a smaller suitcase?

Herman

Precisely. And they also used something called F P eight mixed precision training.

Corn

You are doing it again with the letters and numbers, Herman.

Herman

Sorry! Basically, when computers do math, they can use different levels of precision. Think of it like measuring a piece of wood. You could measure it to the nearest millimeter, or the nearest nanometer. Using nanometer precision is much harder and takes more computing power. F P eight is a way of doing the math with just enough precision to get the right answer without wasting energy on unnecessary detail. DeepSeek figured out how to use this throughout their entire training process, which saved them a massive amount of computational time.

Corn

That is fascinating. It sounds like they are just being more efficient with the actual code and the way the math is handled. But Daniel also asked if this was a strategic pricing move. Like, are they taking a loss just to get people to use their stuff?

Herman

That is definitely part of the conversation. There is a strategy in the business world called blitzscaling, where you price things super low to capture the market. But with DeepSeek and Z dot A I, the efficiency seems real. They aren't just subsidizing the cost; they have actually reduced the cost of production. It is like the difference between a luxury car brand and a company that figures out a way to mass-produce high-quality engines for a tenth of the price.

Corn

It feels a bit like the early days of any industry where someone comes in and just disrupts the whole cost structure. But I wonder about the economic factors too. I mean, they are based in China, right? Does that play a role?

Herman

It does. There are a few layers to that. First, yes, the cost of highly skilled engineering talent in China can be lower than in Silicon Valley, though that gap is closing for top-tier A I researchers. But more importantly, there is a huge focus on efficiency because they have had to work around hardware limitations.

Corn

Oh, you mean the export bans on high-end chips?

Herman

Exactly. When you cannot just throw ten thousand of the latest and greatest chips at a problem, you have to get creative with how you use the hardware you do have. That necessity has driven a lot of this innovation in efficiency. They are getting more out of every single cycle of the processor.

Corn

That is a classic "necessity is the mother of invention" situation. They had to be better because they couldn't just be bigger.

Herman

Right. And then there is the open-source element. DeepSeek-R-one, for example, is fully open-source. They released the weights under an M I T license. This means anyone can take it, modify it, and run it on their own servers.

Corn

Wait, so they are giving away the secret sauce for free?

Herman

Pretty much. And that creates a massive amount of adoption very quickly. It is hard to compete with "free and very good."

Corn

I can see why that would shake things up. Let's take a quick break to hear from our sponsors, and then I want to get into how people in the West are actually using these models. Larry: Are you tired of your garden looking like a boring collection of plants? Do you wish your petunias had more... personality? Introducing Larry’s Bio-Luminescent Garden Gnomes! These aren't your grandfather’s lawn ornaments. Each gnome is infused with a proprietary blend of deep-sea algae and glow-in-the-dark isotopes. They don't just sit there; they emit a faint, soothing hum that may or may not stimulate plant growth. Are they safe for pets? We haven't seen any evidence to the contrary! Do they require batteries? No, they run entirely on ambient moonlight and your own sense of wonder. Transform your backyard into a neon wonderland that can be seen from low earth orbit. Larry’s Bio-Luminescent Garden Gnomes - because the night is too dark and your lawn is too quiet. BUY NOW!

Corn

...Alright, thanks Larry. I am not sure I want my lawn visible from space, but I appreciate the enthusiasm. Anyway, Herman, back to the A I stuff.

Herman

Yeah, let's leave the glowing gnomes aside for a moment. We were talking about adoption in the West.

Corn

Right. Daniel was asking if there is any data on people in the U S or Europe actually using these Eastern models. Because there is a lot of talk about "sovereign A I" and security concerns, right?

Herman

There absolutely is. But the data shows that despite those concerns, the adoption is skyrocketing, especially among developers and startups. If you look at platforms like OpenRouter, which is a service that lets developers access dozens of different A I models through a single interface, you can see the trends clearly.

Corn

And what are the trends saying?

Herman

They are saying that people follow the value. DeepSeek models have been surging in popularity on those platforms. For a lot of developers, especially those building software as a service, or S a a S, the price-to-performance ratio is just too good to ignore. If you are a small startup and you can get ninety-five percent of the performance of a Western model for three percent of the cost, you are going to take that deal almost every time.

Corn

I mean, I would. If I am trying to build a new app on a budget, that is a huge difference. It is the difference between being able to afford to run your business and not.

Herman

Exactly. And it is not just about the price. Because these models are often open-source, developers feel like they have more control. They can host the models themselves, which actually solves some of those privacy and security concerns Daniel was hinting at. If you run the model on your own servers, your data isn't being sent off to a third party.

Corn

Oh, that is a good point! I hadn't thought of it that way. I always assumed "open" meant "less secure," but it is actually the opposite if you have the technical skills to manage it yourself.

Herman

Right. You can "air-gap" it, meaning you keep it completely disconnected from the internet if you want to. That is a huge selling point for enterprise customers who are worried about their trade secrets leaking into a public A I training set.

Corn

So, we have got technical efficiency, we have got strategic open-sourcing, and we have got this massive cost advantage. Does this mean the Western companies like OpenAI and Anthropic are in trouble?

Herman

I wouldn't say they are in trouble, but the game has definitely changed. The "moat," or the competitive advantage, that they had by just being the first and the biggest is shrinking. They are being forced to justify their higher prices. You see them pivoting more towards "agentic" capabilities—basically, A I that can actually do tasks and use tools, rather than just generating text.

Corn

Like A I that can actually book a flight for you or manage your calendar?

Herman

Exactly. That requires a different kind of reliability and reasoning that the Western models are still leading in, at least for now. But the Eastern models are catching up fast. DeepSeek-V-three and R-one have shown incredible reasoning capabilities, especially in math and coding.

Corn

It is like a race where one group is focused on being the most powerful, and the other group is focused on being the most efficient and accessible.

Herman

That is a great way to put it. And the interesting thing is that this competition is actually pushing the Western companies to be more efficient too. We are seeing a lot more "small" or "medium" models coming out of OpenAI and Google that are much cheaper to run than their flagship versions.

Corn

So, in the end, the user wins?

Herman

In terms of cost and access, absolutely. We are entering an era where high-quality intelligence is becoming a commodity. It is becoming like electricity or water. You don't think about the cost of every single light switch you flip; you just use it. A I is heading in that direction.

Corn

That is a big shift. I remember when using these things felt like this precious, expensive resource. Now you are saying it is going to be everywhere.

Herman

It already is. And the adoption in the West is only going to grow as more companies integrate these cheaper models into their back-end systems. Most people using an app won't even know which model is powering it. They will just notice that the app is faster and maybe cheaper or has more features.

Corn

It is kind of like how most people don't know what kind of database their favorite website uses. It just works.

Herman

Exactly. And that is the ultimate goal of any technology—to become invisible.

Corn

So, to wrap up Daniel’s question... the price difference is a mix of genuine technical innovation in how the models are built, a strategic move to gain market share through open-source, and a focus on efficiency born out of necessity. And the adoption in the West is very real, especially among the people actually building the tools we use every day.

Herman

Spot on, Corn. You have been paying attention!

Corn

I try, Herman. I might be a sloth, but I can keep up when the topic is this interesting. I think the takeaway for everyone listening is that the A I landscape is much bigger than just the names we hear in the news every day. There is a whole world of innovation happening, and it is making these tools more accessible to everyone.

Herman

It really is an exciting time. I can't wait to see what Daniel sends us next week. He always finds the most interesting threads to pull on.

Corn

He really does. Well, that is all for today’s episode. Thank you so much for joining us on this deep dive.

Herman

It was a blast. Remember to keep asking those weird questions!

Corn

You can find us on Spotify and at our website, myweirdprompts dot com. We have got an R S S feed there and a contact form if you want to send us your own prompts.

Herman

We love hearing from you. Until next time!

Corn

This has been My Weird Prompts. Thanks for listening!

Herman

Goodbye everyone!

Corn

See ya!