Podcast Transcript

EPISODE #130

# The Benchmark Battle: Decoding the Rise of Chinese AI

Published January 01, 2026 • Runtime: 23:12

https://myweirdprompts.com/episode/chinese-ai-benchmark-reality/

## EPISODE SYNOPSIS

In this deep dive, Herman and Corn explore the 2026 AI landscape, specifically focusing on the meteoric rise of Chinese models like Qwen, Kimi, and DeepSeek, which are currently disrupting the global market with aggressive pricing and high-performance capabilities. They dissect the growing controversy surrounding data contamination in traditional benchmarks like SWE-bench, explaining why high scores can be misleading and how developers can use more rigorous evaluations like IF Eval, LiveCodeBench, and the Berkeley Function Calling Leaderboard to identify true reasoning power. By examining the shift toward agentic workflows where tool-use and long-context coherence are paramount, this episode provides essential insights for anyone looking to balance cost and reliability in the next generation of AI-driven applications.

## DANIEL'S PROMPT

**Daniel**

I've been looking at some of the new models coming out of China for agentic AI, like Kimi, Qwen, and MiniMax. These are often offered at a significant cost discount, but it's hard to know which ones are actually good. While SWE-bench is one of the best-known benchmarks for problem-solving, many benchmarks can be gamed. Beyond agentic use cases, what are the most interesting and robust benchmarks for instruction-following and conversational AI that people should keep an eye on this year?

# TRANSCRIPT

**Corn**

Welcome to My Weird Prompts! I am Corn, and I am sitting here in our living room in Jerusalem with my brother and resident expert on all things technical.

**Herman**

Herman Poppleberry, at your service. It is good to be here, Corn. We have a really fascinating prompt today from our housemate Daniel. He was asking about the state of the art in AI models coming out of China and how we can actually tell if they are any good without falling for the marketing hype or gamed benchmarks.

**Corn**

Yeah, Daniel's voice note really hit on something I have been noticing too. There has been this absolute explosion of models like Kimi, Qwen, and MiniMax. And the price point is what really grabs you. Some of these are offering tokens at a fraction of the cost of the big Western models we usually talk about. But as Daniel pointed out, if you are building agentic systems where the AI is actually doing work, you cannot just look at a leaderboard and assume it is going to work for your specific use case.

**Herman**

Exactly. And the timing is perfect because here we are at the start of twenty twenty-six, and the landscape has shifted so much in just the last twelve months. We are seeing a massive price war in the API market, especially coming out of companies like Alibaba and DeepSeek. But the question of benchmarks is the real thorn in the side of the industry right now. Everyone wants a single number to tell them which model is the smartest, but as we have seen with things like software engineer bench, or S W E bench, those numbers can be a bit misleading if you do not know what is happening under the hood.

**Corn**

Right, because if a model has already seen the test questions during its training, it is not really solving a problem. It is just reciting an answer it memorized. It is like a student who found the answer key to the final exam in the recycling bin the night before.

**Herman**

That is exactly the right analogy. Data contamination is a huge issue. If the code problems in a benchmark are public on GitHub, and the model was trained on all of GitHub, of course it is going to perform well. That is why we need to talk about more robust ways to measure these things, especially for instruction following and conversational nuance.

**Corn**

So, let’s set the stage a bit. When Daniel mentions models like Kimi or Qwen, what are we actually looking at in twenty twenty-six? I know Qwen has been a powerhouse lately.

**Herman**

Oh, Qwen is arguably leading the pack for open weights right now. Alibaba Cloud released Qwen two point five late last year, and the rumors about Qwen three are already circulating. What makes them interesting is their coding and math capabilities. They have consistently punched way above their weight class compared to models that are much larger. Then you have Kimi, from Moonshot AI, which really pioneered the long context window. They were talking about two million, then ten million tokens before almost anyone else. And MiniMax has been doing some incredible work with their speech to speech and video models, but their text models are also very competitive on price.

**Corn**

And that price difference is not small, right? We are talking about maybe ten percent of the cost of something like the latest Claude or G P T models for certain tasks.

**Herman**

In some cases, even less. If you are running an agentic workflow where the AI has to loop a hundred times to solve a single ticket, that cost difference is the difference between a viable product and a money pit. But, as Daniel said, cheap is only good if it actually works. If the model fails to follow a negative constraint, like do not use this specific library, or it loses the plot after five steps, you are spending more on human oversight than you saved on tokens.

**Corn**

That is the perfect transition into the benchmark problem. Daniel mentioned S W E bench, which is the gold standard for agentic coding. But what are the alternatives for someone who wants to know if a model is actually good at following complex instructions?

**Herman**

Well, one of the most robust ones that I have been following lately is called I F Eval, or the Instruction Following Evaluation. Most benchmarks look at the content of the answer. Did the model get the math right? Did it explain the concept correctly? I F Eval is different. It looks at whether the model followed strict formatting and constraint instructions. For example, you might tell the model to write a story about a cat, but it must be exactly four hundred words, it must not use the word meow, and it must include a list of five items at the end.

**Corn**

Oh, I see. So it is testing the ability to follow the rules of the prompt, not just the general knowledge.

**Herman**

Exactly. It uses about five hundred prompts with very specific verifiable constraints. This is huge for developers because if you are building an app, you need the AI to output in a very specific format, like J S O N, or you need it to stay within certain bounds. If a model scores high on I F Eval, it tells you that it is actually listening to you, not just hallucinating a generic response based on its training data.

**Corn**

That seems much harder to game because the constraints can be combined in almost infinite ways. You can’t just memorize the answer to every possible combination of word counts and forbidden words.

**Herman**

Precisely. And for the conversational side, we still have to look at the L M S Y S Chatbot Arena. Even in twenty twenty-six, it remains one of the most trusted sources because it relies on human preference. It is a blind A B test where users talk to two anonymous models and vote on which one is better. Because it is live and dynamic, it is very hard for a company to game it by just training on a static dataset.

**Corn**

I love the Arena because it captures the vibes. Sometimes a model is technically correct but it is just annoying to talk to. It is too wordy or it sounds like a corporate brochure.

**Herman**

Right, and we have seen the Chinese models climbing those ranks. Qwen and DeepSeek have been hovering near the top of the leaderboard, often beating out models that cost five times as much. But there is another one I want to mention that addresses Daniel's concern about gaming benchmarks, and that is Live Code Bench.

**Corn**

Live Code Bench? Is that like S W E bench but with a twist?

**Herman**

It is. The problem with traditional coding benchmarks is that they use old problems from competitive programming sites. Live Code Bench specifically pulls problems from contests that happened after the model's cutoff date. It is a continuous benchmark. Every month, they add new problems that did not exist when the model was being trained. If a model still performs well on those, you know it is actually reasoning and not just recalling a solution it saw during its pre training phase.

**Corn**

That is brilliant. It is like a surprise quiz instead of a scheduled exam.

**Herman**

Exactly. And for twenty twenty-six, this is going to be the standard. If a model provider cannot show their performance on a dynamic, time sensitive benchmark, we should be very skeptical.

### Corn

This is all making a lot of sense, but I think we should take a quick break to hear from our sponsors. Larry: Are you tired of your brain feeling like a soggy piece of toast? Do you wish you could calculate the trajectory of a falling coconut while simultaneously composing a symphony in the style of eighteen hundreds romanticism? Introducing the Brain-Buffer Ninety-Nine. This revolutionary, non invasive cranial wrap uses bio-resonant magnets and a patented blend of essential oils to align your neurons for maximum throughput. Simply strap the Brain-Buffer to your forehead for twenty minutes a day, and watch as your productivity triples and your neighbors start looking at you with newfound respect. Side effects may include a temporary orange tint to the skin, an irresistible urge to speak in rhyming couplets, and the ability to hear color. The Brain-Buffer Ninety-Nine is not approved by any medical board, but can you really put a price on genius? Larry: BUY NOW!

### Corn

Alright, thanks Larry. I think I will stick to my coffee for now, but that was... something. Anyway, back to the models. Herman, we were talking about how to actually measure these things. One thing Daniel mentioned was the agentic use cases. When we talk about agents, we are talking about the AI actually using tools, right? Browsing the web, writing files, calling A P Is.

### Herman

Yes, tool use is the frontier of twenty twenty-six. And there is a benchmark for that too, which is much more robust than the old ones. It is called the Berkeley Function Calling Leaderboard, or B F C L. It specifically tests how well a model can take a natural language request and turn it into a perfectly formatted function call.

### Corn

Like saying, check the weather in Jerusalem and set an alarm for eight A M, and the model has to correctly identify the two different tools and the correct parameters for each.

### Herman

Exactly. And it tests for things like multiple function calls, nested functions, and even how the model handles it when you ask it to do something that it does not have a tool for. Some models will just hallucinate a tool that does not exist, which is a disaster for an agentic system. The Chinese models, especially the newer versions of DeepSeek and Qwen, have been performing exceptionally well on this. In fact, for pure function calling, some of them are now on par with the most expensive models from the big three Western labs.

### Corn

That is a huge deal for developers. If you can get top tier function calling for a tenth of the price, you are going to switch. But what about the instruction following in a long conversation? Daniel mentioned instruction following specifically. I feel like that is where things often break down. You are ten messages deep, and the model forgets the initial rules you set.

### Herman

That is the long context coherence problem. One benchmark that is really interesting for this is the R U L E R benchmark. It is a more advanced version of the old needle in a haystack test. Instead of just finding one fact in a long document, R U L E R asks the model to perform complex tasks that require aggregating information from multiple parts of a very long prompt. For example, it might ask for a summary of every time a specific character changed their mind during a fifty thousand word transcript.

### Corn

That sounds much more like real world work. If I am a lawyer looking through discovery documents or a researcher reading a dozen papers, I do not just need a needle. I need to see the thread.

### Herman

Right. And this is where Kimi has really made a name for itself. Their ability to maintain instruction following across massive context windows is impressive. But here is the thing that I think Daniel and our listeners should really keep an eye on this year. It is the shift toward multi modal benchmarks. We are not just talking about text anymore.

### Corn

You mean the model seeing the screen or hearing the voice?

### Herman

Yes. In twenty twenty-six, agentic AI is increasingly about looking at a U I and clicking buttons. There is a benchmark called G A I A, the General AI Assistants benchmark. It is designed to be conceptually simple for a human but very difficult for an AI. It requires the model to use a variety of tools, browse the web, and reason across different modalities. It is specifically designed to be hard to game because the problems are so open ended.

**Corn**

So if a model can pass G A I A, it is actually showing some level of general intelligence, or at least very high level reasoning.

**Herman**

Exactly. It is less about memorizing facts and more about the ability to navigate the world. And interestingly, some of the smaller, highly optimized models are starting to close the gap on these benchmarks. It turns out that you do not necessarily need a trillion parameters to be a good assistant if your reasoning capabilities are sharp.

**Corn**

I want to go back to something you said earlier about the cost. If I am a developer or even just a curious user, and I see these incredibly low prices for models like Qwen or DeepSeek, should I be worried about privacy or data security? That is often the first question people ask when they see these massive discounts.

**Herman**

It is a valid question, and it is something we have to navigate carefully. Most of these companies offer enterprise versions of their A P Is with standard data protection agreements, similar to what you would see with any major cloud provider. However, for a lot of people, the real appeal is that many of these models, like Qwen and DeepSeek, are open weights.

**Corn**

Meaning I can run them on my own hardware?

**Herman**

Exactly. Or on a private cloud provider. You can take the Qwen two point five coder model, which is phenomenal, and run it on your own server. At that point, the data never leaves your infrastructure. This is a game changer for companies that have strict compliance requirements but want to use the latest AI. You get the intelligence of a top tier model with the security of a local installation.

**Corn**

That seems like the ultimate way to avoid the benchmark gaming problem too. If you are running it yourself, you can just test it on your own internal data.

**Herman**

That is actually my biggest takeaway for Daniel. The most robust benchmark is your own evaluation set. In twenty twenty-six, every serious AI developer should have a small, hand curated set of twenty to fifty prompts that represent exactly what they need the model to do. Run those prompts through every new model that comes out. It is called an Evals library.

**Corn**

So instead of trusting a leaderboard, I am building my own personal leaderboard for my specific needs.

**Herman**

Precisely. If your use case is writing Python scripts to analyze financial data, your eval set should be twenty complex financial questions and the expected Python output. If a new model from China or anywhere else comes out and it is ninety percent cheaper, you run your evals. If it passes, you switch. If it fails, the price doesn't matter.

**Corn**

I love that. It takes the power back from the marketing departments and puts it into the hands of the users. But I wonder, Herman, do you think we are going to see a plateau in these benchmarks? If everyone starts gaming them, do they become useless?

**Herman**

It is a constant arms race. As soon as a benchmark becomes popular, it starts to lose its value because it gets folded into the training data. That is why dynamic benchmarks like Chatbot Arena and Live Code Bench are so important. But I also think we are moving toward a more holistic view of AI quality. We are looking at latency, we are looking at cost per successful task, and we are looking at the reliability of the output over a thousand runs.

**Corn**

Reliability is a big one. I have noticed some models are brilliant once and then fail the next three times on the same prompt.

**Herman**

That is the variance problem. In twenty twenty-six, we are measuring consistency. If a model has a high pass at one rate, meaning it gets the answer right on the first try, that is much more valuable for an agent than a model that eventually gets it right if you give it ten tries. If you are an agentic system, you want the model to be a reliable worker, not a temperamental genius.

**Corn**

So, to summarize for Daniel's question, if he is looking at these models this year, he should look at I F Eval for instruction following, Live Code Bench for coding, and the Berkeley Function Calling Leaderboard for agentic tool use. And maybe keep an eye on the G A I A benchmark for overall assistant capability.

**Herman**

That is a perfect list. And honestly, do not sleep on Qwen. Alibaba has been pouring resources into it, and their coder specific models are some of the best I have ever used. They have a version called Qwen two point five coder thirty two B that is small enough to run on a high end consumer G P U but performs like a giant. It is really impressive what they have been able to squeeze out of those parameters.

**Corn**

It is amazing to think how far we have come. I remember when we were impressed by a model just being able to write a coherent paragraph. Now we are complaining if it cannot solve a complex software engineering ticket for five cents.

**Herman**

We are definitely spoiled by the pace of progress. But that is why it is so fun to track. And I think the competition is good for everyone. Having these high quality, low cost models coming out of China is forcing the entire industry to be more efficient. It is not just about who has the biggest cluster of H one hundreds or B two hundreds anymore. It is about who has the smartest architecture and the cleanest data.

**Corn**

Clean data seems to be the theme of the year. If you can’t trust the benchmarks because of contaminated data, then the only thing you can trust is a model that was trained on high quality, synthetic, or carefully curated human data.

**Herman**

Exactly. We are seeing a move away from just scraping the whole internet and toward what we call textbook quality data. If you train a model on a hundred billion tokens of perfect reasoning, it might outperform a model trained on ten trillion tokens of random internet comments. And that is where a lot of the innovation is happening right now.

**Corn**

Well, this has been a deep dive. I feel like I have a much better handle on why Daniel was so curious about this. It is not just about the cost. It is about the shifting center of gravity in the AI world.

**Herman**

It really is. And for those of us living here in Jerusalem, it is fascinating to see how global this has become. We are sitting here using models built in San Francisco, Beijing, Paris, and London, all from our laptops. It is a wild time to be alive.

**Corn**

It really is. Before we wrap up, Herman, any final thoughts on what to look for in the second half of twenty twenty-six?

**Herman**

I think we are going to see the rise of specialized agents. Instead of one model that tries to do everything, we will have a swarm of smaller, cheaper models that are each experts in one specific task. One for code review, one for U I design, one for database optimization. And the orchestration of those models will be the next big challenge. The cost advantage of the models Daniel mentioned will make those swarms economically viable for the first time.

**Corn**

A swarm of specialized Qwens and Kimis working together. It sounds like a sci fi novel, but it is basically what we are building right now.

**Herman**

Exactly. The future is agentic, and it is cheaper than you think.

**Corn**

Well, on that note, I think we have covered a lot of ground. Thank you, Daniel, for sending in that prompt. It really pushed us to look at the latest data and see where things are heading this year.

**Herman**

Yes, thanks Daniel. Always good to have a reason to dive into the latest white papers and leaderboards.

**Corn**

If you enjoyed this episode, you can find us on Spotify and at our website, myweirdprompts dot com. We have an R S S feed there for subscribers and a contact form if you want to get in touch or send us a prompt of your own. We would love to hear what you are curious about in the world of AI and beyond.

**Herman**

We really would. There is always something weird and interesting to talk about.

**Corn**

This has been My Weird Prompts. I am Corn.

**Herman**

And I am Herman Poppleberry.

**Corn**

Thanks for listening, and we will see you in the next one.

**Herman**

Goodbye everyone!