MY WEIRD PROMPTS

Podcast Transcript

EPISODE #105

Beyond Math Puzzles: The Truth About AI Benchmarks

Published December 26, 2025 • Runtime: 22:25

https://myweirdprompts.com/episode/ai-coding-benchmarks-truth/

EPISODE SYNOPSIS

In this episode of My Weird Prompts, Herman and Corn tackle the growing controversy surrounding artificial intelligence benchmarks. As new models like Claude 4.5 and GLM 4.7 dominate headlines with record-breaking scores, the duo explores whether high performance on math puzzles actually translates to real-world coding productivity. They break down the dangers of data contamination, the rise of "benchmark gaming," and why the industry is shifting toward more rigorous, live testing environments. From the software engineering challenges of SWE-bench to the "surprise quiz" nature of LiveBench, this episode provides a vital guide for anyone trying to separate marketing hype from actual machine reasoning.

DANIEL'S PROMPT

Daniel

There has been a lot of interest lately in AI industry benchmarks, especially when new models report amazing performance. I’m interested in how these models fare for code generation and editing. There’s a lot of skepticism about benchmarks, suggesting that manufacturers might be targeting benchmark performance rather than making the most useful models. Many benchmarks also focus on complex mathematical puzzles, which isn’t how most people use conversational models. Why is there such a focus on mathematical tasks, and which benchmarks would you recommend that are objective, free from vendor bias, and provide a good snapshot of a tool’s capabilities?

TRANSCRIPT

Corn

Hey everyone, welcome back to My Weird Prompts! I am Corn, and I am so glad you are hanging out with us today in our little corner of Jerusalem. As always, I am joined by my brother and resident expert on just about everything.

Herman

Herman Poppleberry, at your service. It is good to be here, Corn. Although, I have to say, the weather today makes me want to just stay inside and read research papers all day.

Corn

Well, that is perfect, because our housemate Daniel sent us a prompt that is right up your alley. It is all about the world of artificial intelligence benchmarks. And honestly, it is something I have been seeing all over my feed lately, but I feel like I am only catching the surface level of it.

Herman

It is a massive topic right now, especially as we wrap up twenty twenty-five. The landscape of AI has changed so much just in the last twelve months, and the way we measure these models is becoming a bit of a controversial subject. Being a donkey, I can be a bit stubborn about wanting the actual data instead of just the marketing hype, so I have been digging into this quite a bit.

Corn

I remember you mentioned you were looking at some new models recently. Daniel specifically mentioned Claude Opus four point five and this new GLM four point seven model from Z dot A-I. He was asking about how these things actually perform when it comes to coding, not just solving math puzzles.

Herman

That is such a great point from Daniel. And it touches on a real frustration in the industry right now. We are seeing these huge announcements where a company says, our new model scored ninety-eight percent on this specific test! And everyone cheers, but then people actually try to use it to write software and they find it still makes the same silly mistakes.

Corn

It is like that kid in school who memorizes the practice test but does not actually understand the subject, right?

Herman

Exactly! That is the perfect analogy, Corn. You are a natural at this, even if you are a bit of a slow-moving sloth sometimes. But seriously, that is what we call data contamination or benchmark gaming. If the questions from the benchmark are included in the AI's training data, which is basically the entire internet, then the AI is not reasoning through the problem. It is just remembering the answer it saw during training.

Corn

So, it is basically cheating?

Herman

In a way, yes. Though the developers might not always be doing it on purpose. When you are scraping billions of pages of data, it is hard to make sure you have not accidentally sucked in the answers to the most popular tests. But Daniel's point about the focus on math is really interesting. Why do you think they focus so much on these complex mathematical puzzles, Corn?

Corn

I mean, I guess because math has a right or wrong answer? It is not like asking it to write a poem where everyone has a different opinion.

Herman

That is a big part of it. Math is objective. It is easy to grade. You can run a script that checks if the model got the number forty-two or not. But there is another reason. Mathematics is often seen as a proxy for raw reasoning capability. The idea is that if a model can solve a high-level calculus problem or a complex logic puzzle, it must be smart enough to handle other things, like coding or legal analysis.

Corn

But is that actually true? Does being good at math puzzles make you a good programmer?

Herman

Not necessarily. Coding is about more than just logic. It is about understanding context, following style guidelines, and managing how different parts of a large system interact. A model might be a genius at a isolated math problem but completely fall apart when you ask it to edit a three-thousand-line script without breaking the existing features.

Corn

That is what Daniel was getting at with the skepticism. He mentioned that manufacturers might be targeting benchmark performance rather than making the most useful models. Is that actually happening in twenty twenty-five?

Herman

Oh, absolutely. There is a lot of pressure to be at the top of the leaderboards. It helps with funding, it helps with sales, and it creates headlines. But we are starting to see a pushback. There was a report recently called The State of AI Coding twenty twenty-five that looked at actual productivity gains instead of just test scores. They found that the median size of a pull request—that is basically a set of code changes—increased by thirty-three percent between March and November of this year. It went from about fifty-seven lines to seventy-six lines.

Corn

Wait, so people are actually getting more done?

Herman

They are. But the interesting thing is that the models that score the highest on the math-heavy benchmarks are not always the ones driving that productivity. Some models are better at what we call "in-context" work. That means they are better at looking at your specific project and understanding how you personally write code, rather than just knowing the general rules of Python or Java.

Corn

That makes sense. I would rather have a tool that knows my specific mess than a tool that knows every math formula but has no idea what I am trying to build. But what about those specific models Daniel mentioned? Claude Opus four point five and GLM four point seven. What is the deal there?

Herman

Well, Claude has been a favorite for a long time because it tends to be very careful. It does not hallucinate as much as some other models. But then you have these new challengers like GLM four point seven. The big story there is often the price-to-performance ratio. You might get ninety-five percent of the capability of a top-tier model for a fraction of the cost.

Corn

I love a good bargain. But I worry that if I go with the cheaper one, it is going to break my website.

Herman

And that is why we need better benchmarks! Daniel asked for recommendations for benchmarks that are objective and free from vendor bias. And honestly, the old ones are starting to fail us. But there are a few new ones that I really trust right now.

Corn

Before we get into the nitty-gritty of those benchmarks, I think we should take a quick break for our sponsors. Larry: Are you tired of your garden looking like a boring collection of plants? Do you want your backyard to reflect the true mystery of the universe? Introducing Chronos-Seeds! These are not your grandmother's petunias. Chronos-Seeds have been exposed to high-frequency tachyon bursts in a basement in New Jersey. We cannot guarantee what will grow, but we can guarantee it will be "interesting." Some customers report flowers that bloom yesterday. Others report vines that hum in a language that sounds suspiciously like ancient Aramaic. Do they need water? Maybe. Do they need your secret thoughts whispered to them at midnight? Definitely. Chronos-Seeds - because the linear flow of time is just a suggestion. BUY NOW!

Corn

...Alright, thanks Larry. I think I will stick to my regular tomatoes, thank you very much. Anyway, Herman, back to the world of AI benchmarks that actually mean something.

Herman

Right. So, if we are looking for things that are hard to "game" and actually represent real-world coding ability, I have a few favorites. The first one I want to mention is called LiveBench.

Corn

LiveBench? Like, it is happening live?

Herman

Exactly! That is the whole point. One of the biggest problems with benchmarks is that they become stagnant. Once a test is released, it is only a matter of weeks before it ends up in the training data for the next AI model. LiveBench tries to solve this by constantly releasing new problems that are based on very recent information—stuff that happened after the models were already trained.

Corn

Oh, that is clever. It is like a surprise quiz that changes every day so you cannot memorize the answers.

Herman

Precisially. It is designed with test set contamination in mind. They use objective evaluation, meaning they have very strict ways of measuring if the answer is correct, but the problems are fresh. If a model scores well on LiveBench, it is a much better indicator that it can actually think on its feet.

Corn

Okay, that sounds great for general intelligence. But what about the coding stuff Daniel was asking about?

Herman

For coding, the gold standard right now is something called SWE-bench. That stands for Software Engineering Benchmark. Instead of asking the AI to solve a tiny puzzle, SWE-bench gives it a real-world issue from a popular open-source project on GitHub.

Corn

Like a real bug that a human had to fix?

Herman

Yes. The AI is given the entire codebase, a description of the bug, and it has to actually write the code to fix it. Then, the benchmark runs the project's existing tests to see if the AI actually fixed the problem without breaking anything else.

Corn

That sounds incredibly hard.

Herman

It is! For a long time, even the best models were scoring less than ten percent on this. But in late twenty twenty-four and throughout twenty twenty-five, we have seen those numbers start to climb. Anthropic's Claude models and Google's Gemini models have been doing some really impressive work here. It is a much better reflection of what a software engineer actually does all day. It is not just writing code; it is navigating a complex system.

Corn

So if I am a developer and I want to know which model to use, I should look at the SWE-bench scores?

Herman

That is one of the best places to look. Another one that Daniel might find useful is the Aider leaderboard. Aider is a popular tool that people use to code with AI inside their actual projects. They maintain a leaderboard that specifically tests how well models can perform "refactoring" and "editing" tasks.

Corn

Refactoring... that is just a fancy word for cleaning up code, right?

Herman

You got it. It is about changing the structure of the code without changing what it does. It is one of the most common tasks for a programmer, and it is something that models often struggle with. They might fix the thing you asked for, but accidentally delete three other things in the process. The Aider leaderboard is great because it is based on real-world usage of these models in a very popular coding tool. It is much harder to "game" because it is testing the model's ability to follow complex editing instructions.

Corn

That sounds way more useful than a math puzzle. Why are we still even talking about the math puzzles then?

Herman

Well, to be fair to the math puzzles, they do show us something about the "ceiling" of a model's logic. If a model cannot solve a high-school level math problem, it is probably going to struggle with complex logic in a program. But you are right, we are seeing a shift. There is a growing sentiment that "benchmarking is broken" if we only focus on those static, academic tests.

Corn

I saw a headline about that. It said something about AI reviewing AI? That sounds like a recipe for a disaster.

Herman

It can be! That is another trend in twenty twenty-five. Because these tasks are getting so complex, humans sometimes have a hard time grading them quickly. So developers use a "stronger" AI to grade the "weaker" AI's work.

Corn

Wait, so the teacher and the student are both robots?

Herman

Exactly. And you can see the problem there. If the teacher-AI has the same biases or the same gaps in knowledge as the student-AI, it might give it a passing grade even if the answer is wrong. There was a case study recently about "Oracle validity"—basically, how do we know the person or thing giving the answer actually knows the truth? In coding, we are lucky because we can run the code and see if it works. In other fields, it is much harder.

Corn

This is all making me a bit skeptical of any number I see now. If Daniel is looking at a new model like GLM four point seven and he sees a high score, what should his first question be?

Herman

His first question should be: "Was this an open-ended benchmark or a closed one?" And then: "How does it perform on SWE-bench or LiveBench?" If a company only reports scores on something like HumanEval—which is a very old and very famous benchmark—you should be a bit suspicious.

Corn

Why HumanEval?

Herman

HumanEval was released years ago. It is a collection of about one hundred and sixty-four coding problems. Because it is so old and so famous, it is almost certain that every single one of those problems is in the training data of every new model. It is like taking a history test when you already have the answer key in your pocket. It does not prove you know history; it just proves you can read the answer key.

Corn

So if a model says it got one hundred percent on HumanEval, it is basically just saying "I have read the internet."

Herman

Pretty much! It is still a useful baseline—if a model fails HumanEval, it is definitely not ready for prime time. But a high score there does not mean it is a great coding assistant. For that, you really want to see how it handles those larger-scale tasks.

Corn

You mentioned Google’s Gemini earlier. I feel like they were a bit behind for a while, but I have been hearing more about them lately. How are they doing in the coding world at the end of twenty twenty-five?

Herman

Gemini has actually made a huge comeback. They have a massive "context window," which means the model can "remember" or "look at" a huge amount of information at once—sometimes millions of tokens. For a programmer, that means you can feed the AI your entire documentation, your entire codebase, and all your style guides, and ask it a question. It does not have to guess; it can actually see your whole project.

Corn

That sounds like a game changer. I mean, I can barely remember what I had for breakfast, let alone a whole codebase.

Herman

Exactly. And that is a different kind of "smart" than being good at a math puzzle. It is about information retrieval and synthesis. So, when Daniel asks which benchmarks to recommend, I would say look for those that test "long-context" reasoning too.

Corn

Okay, so let me see if I can summarize this, because my sloth brain is starting to get full. We have LiveBench for fresh, non-contaminated problems. We have SWE-bench for real-world software engineering tasks on GitHub. And we have the Aider leaderboard for practical, day-to-day code editing.

Herman

You nailed it, Corn. And I would add one more thing for Daniel: personal benchmarking. In twenty twenty-five, the most sophisticated users are not just trusting the online scores. They have their own little "test set" of problems that they know are hard or specific to their work. When a new model comes out, they run it through their own five or ten problems to see how it handles them.

Corn

That makes so much sense. Like, if I have a specific way I like my Python scripts to look, I should see if the new model can actually follow my style.

Herman

Exactly. No benchmark is going to be as perfect as your own experience. But the industry is moving in the right direction. We are moving away from those "look how smart I am at puzzles" tests toward "look how much work I can actually help you get done" tests.

Corn

It feels like the "hype" phase of AI is finally starting to settle into a "utility" phase.

Herman

I think that is a very astute observation. We are seeing a lot more focus on reliability. People are realizing that a model that is eighty percent accurate but tells you when it is unsure is actually much more useful than a model that is ninety percent accurate but lies to you the other ten percent of the time.

Corn

Oh, I hate it when they lie. It is so confident too! It will give you a piece of code that looks perfect, and then you run it and your computer starts smoking.

Herman

Hopefully not literally smoking! But yes, that "hallucination" problem is why benchmarks like SWE-bench are so important. They require the code to actually pass a test. You cannot just look good; you have to work.

Corn

So, looking forward to twenty twenty-six, do you think we will ever have a "perfect" benchmark? One that everyone agrees on?

Herman

Probably not. As the models get smarter, the benchmarks have to get harder. It is a constant game of cat and mouse. But I do think we will see more "dynamic" benchmarks. Instead of a static list of questions, we might see benchmarks that are generated by other AIs in real-time to test specific weaknesses. It is going to be a very interesting year for AI evaluation.

Corn

Well, I feel a lot better about this now. It is not just about the numbers; it is about what those numbers are actually measuring. Daniel, I hope that helps you navigate the sea of AI announcements. It sounds like Claude and Gemini are still the big players, but keep an eye on those newer models like GLM if they start showing up on the more rigorous leaderboards.

Herman

And don't be afraid to try them out! A lot of these newer models offer free trials or very cheap API access. The best way to see if a model is "gaming" the benchmark is to give it a task it has definitely never seen before—something unique to your own life or your own project.

Corn

That is great advice, Herman. Even for a donkey, you are pretty sharp.

Herman

Hey, I take that as a compliment! Donkeys are very intelligent and hardworking animals, you know.

Corn

I know, I know. And I am just here to keep us relaxed and ask the questions everyone else is thinking. This has been such a great deep dive. I feel like I actually understand why those math puzzles are everywhere, even if I still think they are a bit silly for testing a coding bot.

Herman

They have their place, but they are definitely not the whole story. I am glad we could clear that up. It is always fun to dig into the data and see what is actually happening behind the marketing curtain.

Corn

Absolutely. Well, I think that is all the time we have for today. Thank you so much for joining us for this episode of My Weird Prompts.

Herman

Yes, thank you everyone. And thank you to Daniel for such a thoughtful and timely prompt. It is exactly the kind of thing we love to explore.

Corn

If you have a prompt you want us to tackle, whether it is about AI, history, or why sloths are clearly the superior species, head over to our website at myweirdprompts dot com. We have a contact form there, and you can also find our RSS feed and links to all our episodes on Spotify.

Herman

We love hearing from you, so don't be shy.

Corn

Until next time, stay curious, and maybe don't buy those time-traveling seeds Larry was talking about.

Herman

Definitely don't do that. Goodbye, everyone!

Corn

Bye! This has been My Weird Prompts. See you in the next one!