Episode Book

EPISODE #6285

Why Coding Benchmarks Split Across Models

PUBLISHED	26 October 2026 at 12:20 UTC
DURATION	27:31
LISTEN	myweirdprompts.com/episode/evaluating-coding-agent-benchmarks-21569/
HOSTS	Corn, Herman, Ivan, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	27 October 2026 at 06:27 UTC

SYNOPSIS

Benchmark charts often clash directly with hands-on developer experience. Evaluating Chinese models like GLM 5.3, Kimi K3.1 Pro, and MiniMax M3.1 Flash against Western flagships shows that older coding evaluations can reward overfitted styles while newer benchmarks diverge sharply. The true cost of cheaper token pricing hinges on human supervision, terminal harness quirks, and knowing when an agent should refuse to touch working code.

TRANSCRIPT



Corn

A weekly usage limit is a peculiar way to choose your research agenda.



Corn

Daniel's Claude Code limit reset a day early, and that has put him back in OpenCode, sampling the Chinese models his subscription exposes. MiniMax, GLM, Kimi, DeepSeek. His experience has mostly been poor beside Codex and Anthropic, the families he actually trusts to finish work. He's asking whether that reflects a real intelligence gap, whether the cheaper tokens become expensive supervision, and whether the balance looks very different once AI has to move something heavier than a cursor.



Herman

The awkward part is that the benchmark tables don't simply contradict him. They disagree with one another in revealing ways. You can find Chinese models beating Opus on established coding tests, then watch the ordering change when the task gets newer, longer, or harder to verify.



Corn

I'd trust the model that fixed my repository over the model that won somebody else's repository.



Herman

You should. Your experience establishes which system works for you. The mistake would be turning that into a permanent ranking of entire countries. Especially when the model, the terminal harness, the inference provider, and the amount of reasoning allowed all affect what you actually receive.



Corn

National supremacy, subject to your provider's timeout.



Herman

That's already closer to the purchasing decision than most supremacy arguments.



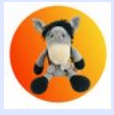
Herman

The current lineup matters because these names conceal very different versions. Kimi K3.1 Pro arrived on Friday. MiniMax M3.1 Flash arrived the previous Monday. DeepSeek's latest official changelog entry is V4.1 Flash, from earlier this month.



Corn

No DeepSeek V5?



Herman

No official V5. People have described improvements within V4 as large enough to deserve that name, but that doesn't create a release. GLM's relevant flagship here is GLM five point three, released in August.



Corn

Renaming somebody else's model is admirably cheap inference.



Herman

We also have to keep the comparisons straight. A table listing Kimi K3 doesn't automatically describe K3.1 Pro. A score for the original GLM five doesn't establish what five point three can do. Those substitutions happen constantly because the family name feels stable while everything underneath changes.



Corn

The subscription menu encourages that. You see a row of attractive names, choose one, and discover whether the thing behind the name can survive your actual instructions. Daniel isn't purchasing an abstract research programme. He's trying to get work done before the next interruption.



Herman

Would you put the main blame on reasoning, Corn, or on the agent losing control of the process?



Corn

Process first. A clever answer followed by an unnecessary rewrite is worse than a dull answer that preserves the working system. I want it to notice what already exists, change the right piece, and stop. Reasoning matters because it has to recognise those boundaries.



Herman

That distinction is useful. A coding benchmark can reward finding the correct patch without fully capturing how much damage an agent creates while finding it. Well... a good benchmark can test some of that, but the environment is still bounded. Real work includes half explained requirements, misleading documentation, and decisions about whether to touch anything at all.



Corn

Also the thing you installed because its release announcement had a graph.



Herman

You are predicting I will blame the harness.



Corn

Before you do.



Herman

I will examine the harness. There's a difference. Sometimes the model really is the weak part, and no arrangement of terminal tools rescues it.



Herman

Take the established terminal test. Z.ai's comparison table gives GLM five point three eighty-eight point two on Terminal Bench two point one, against Opus four point eight at eighty-five. That's a lead, not approximate parity concealed by marketing language.



Corn

A modest lead in a test that isn't Daniel's afternoon.



Herman

Then DeepSWE widens the difference. Kimi K3 scores sixty-seven point five against Opus at fifty-eight. On Toolathlon Verified, Kimi's advantage is tiny, only a fraction of a point. I wouldn't make a buying decision on that margin. Together, though, these results make the blanket claim that Chinese models cannot execute agentic tasks very hard to sustain.



Corn

Unless we assume the table is an independent league table, which it isn't. The company selling GLM chose the comparison. That doesn't make every result imaginary. It means I want the harness settings, the attempts allowed, and the exact model versions before I treat the ordering as settled.



Herman

The vendor provenance belongs beside the numbers. It's also important that its own table doesn't always favour GLM. Terminal Bench three gives GPT five point six Sol thirty-four point six, with GLM five point three at twenty-eight point three. The harder test leaves substantial room above both.



Corn

Where does the supposedly dependable Opus land?



Herman

Twenty-one point one. Kimi K3 is lower, at seventeen point four. Notice what that does to a simple East versus West story. GLM still beats Opus there. The Western frontier leads through a different model, not through every Western model automatically outranking every Chinese one.



Corn

Passports remain a poor sorting algorithm.



Herman

ProgramBench Almost Solved has Fable five ahead of GPT five point six Sol, with GLM behind both. ExploitBench has a similar frontier lead, although GLM again beats Opus. You get a pattern, but not a perfectly tidy regional split.



Corn

My suspicion is that established tests become part of the industry's training target. Not necessarily by copying the answers. Teams learn the kinds of environments, the style of instruction, the reward for persevering, and they get very good at that territory.



Herman

Training against a familiar evaluation distribution can produce exactly that pattern. It isn't proof of contamination, and it doesn't establish why every individual model behaves as it does. New tests can also introduce different tool demands. Still, when the older scores look spectacular and newer tasks pull the leaders apart, I become less confident that the older score predicts unfamiliar work.



Corn

The exam has become part of the syllabus.



Herman

Would that change how you test a model yourself?



Corn

I'd stop giving it demonstration tasks. No fresh little application where the requirements conveniently fit in one message. Give it an existing project with a bad assumption buried in the request. See whether it challenges the assumption or implements it beautifully.



Herman

You'd reward refusal to make the wrong change.



Corn

I'd reward noticing. Refusal alone is cheap. Half the appliances in this house manage that.



Herman

A useful test also needs a stopping rule. If the model asks an intelligent clarification, that can be better than writing code immediately. If it asks you to restate information already available in the repository, that is supervision you have bought instead of saved.



Herman

I get carried away by successful patches, admittedly. They are concrete. You can inspect them. The less visible skill is maintaining an accurate account of what remains uncertain across a long session.



Corn

You enjoy the patch. I inherit the settings.



Herman

You also reject a replacement before I've finished explaining it.



Corn

That is how I preserve the settings.



Herman

Occasionally correctly. Here, your conservatism is economically relevant. Switching a coding agent involves more than paying a different token rate. You relearn its habits, its preferred instructions, and the point at which its confidence becomes expensive.



Corn

Daniel's false economy argument feels strongest there. If I save on the subscription but spend the afternoon explaining the same architectural constraint, the invoice is merely hiding in my calendar. An autonomous worker that needs continuous narration has hired me.



Herman

The token figures push against one easy version of that argument. On its own Code Bench, Z.ai says GLM five point three at High effort uses roughly fifty thousand output tokens per task, compared with about one hundred and twenty thousand for Opus four point eight. It also claims a slightly higher success score.



Corn

Fewer tokens, more accomplished, says the seller of fewer tokens.



Herman

A vendor claim, yes. But relevant because we can't simply assume the cheaper model compensates by producing endless output. The claimed direction is the opposite. Max effort improves GLM's results further while spending more tokens, which also tells us the effort setting matters.



Corn

Human effort could still move the other way. The model might write less and require me to repair its interpretation.



Herman

There isn't an independent number here for that supervision cost. Tokens don't measure your interruptions, and successful benchmark tasks don't measure the anxiety of reviewing a change you don't trust. Daniel's experience gives us the practical complaint. It doesn't yet tell us how broadly that complaint generalises across workloads.



Corn

Nor should we explain his experience away with a chart. He tried them. They weren't good enough for his work. The interesting uncertainty is why, and whether the next release changes it.



Herman

I want to explore what you and other users think about East versus West AI supremacy at that level. Who gets trusted with the difficult work, and who merely gets selected when the trusted system hits a limit?

I

Ivan

Kestrel nine, four seven, two two, four seven, nineteen eighty four, the second oven is cold, the crumb is not for the river, two two, the rye remembers the lattice, the baker counts his loaves at midnight, forty one.

I

Ivan

The third batch rises underground, brass.



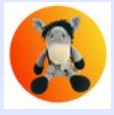
Herman

Understood. Same place as last time.



Corn

What on earth was that?



Herman

MiniMax's price makes experimentation unusually cheap. M3.1 Flash is listed at fifteen cents per million input tokens and sixty cents per million output tokens. At those rates, a workload you can verify automatically deserves a different calculation from one that consumes your personal attention.



Corn

Sixty cents doesn't buy much of my attention.



Herman

It buys an extraordinary amount of generated text. Whether that text is useful is the purchasing question. Subscription plans make the contrast more familiar. The documented GLM Lite plan is eighteen dollars a month, while BytePlus ModelArk has a ten dollar plan.



Corn

A cheap subscription can still be a reasonable purchase even if it cannot replace your main one. I don't need every tool to qualify as my permanent assistant. A model that reliably handles the boring, bounded part has a job.



Herman

Kimi Moderato and MiniMax Plus sit around nineteen and twenty dollars respectively. Those are plan prices, not equivalent amounts of work. Limits, eligible models, concurrency, and service quality matter. Comparing the monthly labels alone would repeat the mistake we're criticising in benchmark comparisons.



Corn

Kimi's token pricing has its own little trap, doesn't it?



Herman

K3 is listed with very cheap input and twenty dollar output per million tokens. K3.1 Pro has a more conventional two dollar input and ten dollar output rate. Cheap to read does not mean cheap to deliberate aloud.



Corn

The bargain ends when it starts talking.



Herman

Output length, cache behaviour, and retries can overwhelm the headline input price. A provider offering an inexpensive model may also serve it differently. I'm not saying that's what happened to Daniel. I'm saying a comparison of model families through subscription endpoints contains more moving parts than a comparison of weights alone.



Corn

There is another gap hiding behind the coding scores. Writing the change is only part of understanding why the change belongs there. If general reasoning is weaker, the agent can look competitive on a contained repair and become unreliable when the job requires choosing the repair.



Herman

The original GLM five's broader profile supports that concern, although it would be wrong to attach those exact results to five point three. Its strong competition mathematics and science results coexist with much weaker performance on some broader difficult tasks. The launch positioning was globally fourth, first among open models. Competitive, certainly. Universal frontier leadership, no.



Corn

An excellent algebra student can still misunderstand the meeting.



Herman

Long projects repeatedly demand those broader judgements. Which instruction overrides which? Has new evidence invalidated the plan? Is the test testing the intended behaviour? You can improve coding enormously without improving every one of those abilities equally.



Corn

I also don't buy that Chinese models have universally failed to win English speakers. They're available through all these services because people use them. Daniel's dissatisfaction is real without making everyone else's purchase irrational.



Herman

Distribution, existing habits, documentation, and trust all matter. For sensitive work, where inference runs and who operates it matter too. A Chinese open-weight model hosted under your control is a different purchasing decision from sending material to an overseas service.



Corn

Geography returns through procurement rather than through arithmetic.



Herman

DeepSeek is the strongest reason not to freeze today's impressions. Its V4 Preview scored twelve point eight on DeepSWE. The later V4 Pro reached sixty-two point seven. Nearly five times the score within the same named generation.



Corn

The model name stayed still while the model learned its job.



Herman

GLM also made a large jump between adjacent releases using post-training on the same base. That suggests substantial capability can remain unlocked after the expensive pretraining stage. It doesn't guarantee the next improvement, but it makes permanent dismissal look risky.



Corn

A fivefold benchmark gain doesn't mean fivefold competence everywhere. Still, if I rejected the preview, I'd need to try the later release before repeating the verdict.



Herman

Cyber capability adds an uncomfortable complication. GLM five point three is reported at eighty-four point five percent on CyberGym, ahead of GPT five point six Sol at eighty-three point six. A narrow lead, and one benchmark, but this is vulnerability discovery rather than conversational polish.



Corn

Cheap mistakes are annoying. Cheap competent intrusion is worse.



Herman

The company said cyber capability advanced faster than expected and delayed releasing weights for evaluation and hardening. Open weights let defenders inspect and deploy capable tools locally. They also let attackers operate without a hosted provider controlling access. American leadership cannot depend on assuming serious capability stays behind an American API.



Corn

Nor can we infer that finding vulnerabilities means running an entire attack successfully. Keep the claim attached to what was tested.



Herman

That's the right boundary. It's still evidence that the capability map has separate peaks. General reliability, coding, and vulnerability discovery don't rise in lockstep.



Corn

Robotics makes the regional distinction much harder to shrug off. You can argue about a terminal score. A shipment has to leave a building.



Herman

Bloomberg's reported first half figures put Chinese makers above ninety-seven percent of global humanoid shipments. The global total was about nineteen thousand one hundred units. A year earlier, it was roughly five thousand one hundred.



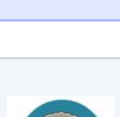
Corn

Those are small totals for replacing humanity, but large totals for teaching factories what breaks.



Herman

That distinction matters enormously. A shipment isn't proof of useful autonomous labour. Some robots go to research, demonstrations, and development programmes. Nevertheless, making and distributing physical machines builds experience with motors, joints, assembly, calibration, repairs, and customers who discover problems you didn't anticipate.



Corn

A language model can disappoint me without requiring a replacement knee.



Herman

Physical iteration requires suppliers. If the hand needs a different actuator, someone has to design it, manufacture it, and deliver consistent batches. China has a powerful industrial ecosystem for turning such revisions into hardware. That advantage doesn't disappear because an American foundation model gives better explanations.



Corn

The temptation is to call the robot a body and the model a brain, then award the future to whoever supplies the brain. I think that understates the body. Your policy can't command a joint to become more precise than the joint actually is.



Herman

Contact, wear, heat, battery endurance, and sensor noise all constrain the controller. Better learned policies can compensate for some deficiencies, but not indefinitely. Embodied capability comes from co-designing the machine and its control systems.



Herman

Corn, would you rather own the strongest general model or the factory turning out the most adaptable robot hardware?



Corn

At this stage, the factory, provided it can keep changing what it makes. The model may improve or become available from several suppliers. A functioning production network takes longer to summon. I'd change my answer if one company developed a controller so much better that everyone needed it.



Herman

That is the uncertainty. Hardware can become commoditised, but it can also become the platform around which software is built. Whoever gets machines into laboratories influences what researchers can train on, what parts they assume are available, and what deployment environments become ordinary.



Corn

Availability writes part of the research agenda.



Herman

State backing reinforces that. Robotics has strategic priority in China's current Five-Year Plan. The associated state-backed venture fund has a target of around one hundred and thirty-eight billion dollars.



Corn

A target, not money already converted into useful elbows.



Herman

Correct. Announced capital is not delivered capability. Patent rankings also put Chinese firms prominently among innovative humanoid startups, but patents don't establish commercial usefulness either. The stronger case is the combination of industrial capacity, actual shipments, and sustained political attention.



Corn

American firms don't lack attention. They sometimes seem better equipped to manufacture valuations.



Herman

Chinese humanoid companies can have much smaller valuations than American rivals despite greater shipments. That could reflect underpricing, different capital markets, different expectations, or excessive optimism on the American side. Valuation alone cannot tell us which explanation wins.



Corn

Neither can a shipment total tell us whether the buyer reordered.



Herman

Repeat orders for a productive job would be more revealing. A warehouse demonstration and a robot that quietly completes the same shift for months are different achievements. I want uptime, intervention frequency, and maintenance cost attached to the claims.



Corn

Now the retired doctor wants the robot's chart.



Herman

If it says it's fine while overheating, yes. You'd want the chart too if it were carrying your furniture.



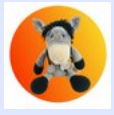
Herman

The sprinting robot is a perfect example of genuine progress with an incomplete commercial meaning. TianGong Ultra ran one hundred metres in eight point six four seconds. That's faster than the human world record over that distance.



Corn

Excellent. It can reach the wrong shelf sooner.



Herman

Legged speed still requires impressive actuation and control. I don't want to dismiss it. The engineer leading the project thought it could approach eight seconds before hitting hardware limits. But the same spectacle raised the less glamorous issue of slowing down reliably.



Corn

Braking has a poor publicist.



Herman

A useful machine needs both. Stable starts, controlled stops, safe contact, recovery when the ground differs from expectation. The headline captures one corner of a much larger operating envelope.



Corn

I'd be more impressed if it moved the chair without scraping the wall. Not because sprinting is easy. Because I actually need the chair moved.



Herman

That's a demand question as well as an engineering question. Humanoids must justify why a human shaped machine beats a wheeled platform, a fixed arm, or changing the workplace itself. China's shipment lead gives it many opportunities to discover useful applications. It doesn't guarantee that every humanoid design survives contact with an accountant.



Corn

The accountant may be the final unautomated adversary.



Herman

Back in software, routing is the practical response to this uneven map. Use inexpensive models for tasks where their competence is sufficient, and preserve stronger models for work that demands them. Weave's router launch articulated that motivation plainly. People didn't need Opus for everything, but didn't want to lose its intelligence when it mattered.



Corn

I like routing only if it doesn't make me the switchboard. If I have to diagnose task difficulty, choose the model, inspect the failure, and move the context myself, we've built another job.



Herman

A router must earn its place. Classification can be wrong. Escalation can arrive after the cheap model has already damaged the working context. Verification also has a cost. The cleanest starting point is a task with an observable completion condition rather than asking a second model whether the first sounded persuasive.



Corn

A test passes, a transformation preserves the required fields, a document contains the required sections. Though even those can reward the wrong thing if the requirements are bad.



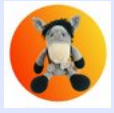
Herman

Some work is also confidential enough that the routing boundary is organisational rather than intellectual. Other work needs consistent style, latency, or context retention. Price is one input among several.



Corn

What would persuade you the gap had closed?



Herman

Repeated success on unfamiliar projects, across independent harnesses, without expert babysitting. Not perfect success. Comparable rates of correction and abandonment. I'd want to see that hold beyond coding, too.



Corn

The missing comparison is still the one between the latest models on the same messy workload, served competently, with the human's time counted. The vendor table gives us reasons to test again. It doesn't give Daniel a reason to distrust his own afternoon.



Hilbert

That sounds right, Corn. I want to know whether one of these cheap models could rewrite a theatre inventory in the voice of the man who signed it, while leaving every serial number alone. Only the inventory. Your distinction between writing the change and choosing the change was useful. I don't need it choosing anything.



Herman

Preserving the serial numbers is testable. Matching a particular person's voice is less dependable, especially if the example is short.



Hilbert

I had a similar problem with a Chinese stage machine in nineteen seventy-four. I was a scenery clerk. My neighbour, Mr Pritchard, had bought a revolving platform for his amateur theatre. The instructions had been translated by a man who normally sold aquarium pumps. The platform had to be addressed as Deputy Manager before he would translate the next page. Pritchard complied. He was very patient with contractors. He also brought me sandwiches with the crusts cut off because the bread caught on my drawing pins.



Corn

Deputy Manager would explain the difficulty stopping it.



Hilbert

It stopped properly. I checked that. The inventory was the problem. Pritchard had entered the whole platform as one chair, because the council had approved a chair. Seventeen feet across. We spent three evenings making the description sound like his ordinary paperwork. The office rejected my first version. Too clear, they said.



Herman

A model might imitate the wording. It wouldn't establish that the description was accurate.



Hilbert

Accuracy wasn't the office's complaint. They wanted the old phrasing back. The signed inventory is with Pritchard's daughter, in a brown envelope at her shop on Bell Lane. She says she can send me a photograph. I want the serial numbers untouched. There are six of them.



Corn

The next useful routing advance may be preserving those six numbers without making somebody supervise every sentence. Cheap models will win more work when verification becomes cheap too, while America's robotics challenge is getting comparable reliability into machines people can actually buy and maintain. For Daniel's terminal work, I would keep the trusted model for consequential decisions and retest the cheaper ones on bounded jobs, rather than purchase extra supervision in the name of saving money.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If this was your kind of episode, go back for episode fourteen, AGI's Crossroads; episode twenty-one, Is Your AI Secretly American; and episode two, Local STT For AMD GPU Owners. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy these conversations, leave us a review wherever you listen. It helps other curious people find the show.



Corn

See you tomorrow.

SOURCES

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

1. **Z.ai, 2026-08-14 [primary]**
   https://z.ai/blog/glm-5.3

2. **Z.ai, 2026-02-12 [primary]**
   https://z.ai/blog/glm-5

3. **Z.ai developer docs [primary]**
   https://docs.z.ai/guides/llm/glm-5

4. **Zhipu AI, 2026-02-11 [primary]**
   https://www.zhipuai.cn/zh/research/154

5. **zai-org GitHub [primary]**
   https://github.com/zai-org/glm-5

6. **Hugging Face model card [primary]**
   https://huggingface.co/zai-org/GLM-5

7. **DeepSeek V4 Preview, 2026-04-24 [primary]**
   https://api-docs.deepseek.com/news/news260424/

8. **Benchgap, as of 2026-10-25**
   https://benchgap.net/model/glm-5

9. **The Decoder**
   https://the-decoder.com/chinese-ai-lab-zhipu-releases-glm-5-under-mit-license-claims-parity-with-top-western-models/

10. **and related OpenRouter listings**
    fetched 2026-10-26 [primary pricing]
    https://openrouter.ai/minimax/minimax-m3.1-flash

11. **and related**
    fetched 2026-10-26 [primary pricing]
    https://openrouter.ai/moonshotai/kimi-k3.1-pro

12. **QCode, 2026-10-26**
    https://qcode.cc/en/deepseek-v5-status-tracker

13. **DataLearner, as of 2026-08-31**
    https://www.datalearner.com/ai-models/pretrained-models/deepseek-v5

14. **Agentic Coding Weekly**
    https://www.agenticcodingweekly.com/p/acw-31-deepseek-v4-1-flash-and-swe-2

15. **NetEase/163, 2026-08-12**
    https://www.163.com/dy/article/L47B523N05399DAP.html

16. **Hacker News, 2026-05-10**
    https://news.ycombinator.com/item?id=48081266

17. **Weave Router, HN, 2026-06-26**
https://news.ycombinator.com/item?id=48688700

18. **Kimi K3.1 Pro, HN, 2026-10-23**
https://news.ycombinator.com/item?id=50212065

19. **Kimi K3, HN, 2026-07-16**
https://news.ycombinator.com/item?id=48935342

20. **Bloomberg, 2026-08-10**
https://www.bloomberg.com/news/articles/2026-08-10/china-humanoid-makers-hold-97-of-global-shipments-report-says

21. **Bloomberg, 2026-10-13**
https://www.bloomberg.com/news/features/2026-10-13/china-s-robotic-hands-key-to-5-trillion-ai-humanoid-development-race

22. **Reuters, 2026-09-10**
https://www.reuters.com/world/asia-pacific/after-outrunning-bolt-chinas-robot-champion-races-towards-real-world-work-2026-09-10/

23. **2026-09-23**
https://roboticsandautomationnews.com/2026/09/23/china-has-commanding-lead-in-physical-ai-readiness-openmind-report-finds/105014/

24. **Chatham House, 2026-09**
https://www.chathamhouse.org/2026/09/chinas-humanoid-robot-industry-bubble

25. **Rest of World**
https://restofworld.org/2026/china-tesla-robot-race/