

# Claude Code and Codex Score Zero Running DeepSeek on kw-bench, OpenCode Gets 0.54

2026-10-28 / 00:19:00

*“Below the frontier, the harness decides how far a model gets.”*

— from this episode's transcript

- Lenar Kess
- Damra Vol

A new benchmark ran one model, DeepSeek v4.1 Flash, through four harnesses for two hours each. Two independent harnesses scored about half marks and the two lab harnesses scored zero. We go through that result and Arize's engineering behind agents, then Genie 4, the agent credential and exfiltration papers, two Anthropic stories on training data, and a few smaller items.

- [kw-bench: a VAT reconstruction across three generations of SAP](#)
- [Arize on Code Mode](#)
- [Arize on how Alex is built](#)
- [Spine: a subtask-tree context runtime](#)
- [AgentBaton: mid-task model handoff](#)
- [DeepMind's Genie 4](#)
- [Two Sigma on agent controls](#)
- [Keycard on agent credentials](#)
- [Log-To-Leak](#)

- [Stanford Daily on the Claude training-data allegation](#)
  - [Latent Space interview with Anthropic](#)
  - [bioRxiv: thirty agents, eighteen pipelines](#)
- 

## SEGMENTS

[00:00:04](#) Same model, same two hours

[00:08:09](#) Genie 4, minus the numbers

[00:09:40](#) Whose credentials does the agent carry?

[00:12:34](#) Anthropic, training data, and thinking blocks

[00:14:48](#) Thirty agents, eighteen pipelines

[00:16:07](#) Smaller items

## Transcript

### 1. Lenar Kess 00:00:04

It's Wednesday, October 28th. Here's an experiment from a new benchmark paper. Take one model, DeepSeek v4.1 Flash, give it the same task and the same two hours, and run it inside four different harnesses. Two of them score about half marks. The other two score zero, and those two are Claude Code and Codex. The paper is called kw-bench, and it's the lead today. After it we've got Arize on how they build agents, Google's Genie 4, and a run of items about agent credentials and what leaks. Then two Anthropic stories on training data, and some smaller things at the end.

- [youtube.com](#)
- [youtube.com](#)
- [arxiv.org](#)
- [arxiv.org](#)
- [arxiv.org](#)
- [blog.google](#)
- [youtube.com](#)
- [deepmind.google](#)

- [youtube.com](https://youtube.com)
- [youtube.com](https://youtube.com)
- [formal.ai](https://formal.ai)
- [arxiv.org](https://arxiv.org)
- [arxiv.org](https://arxiv.org)
- [stanforddaily.com](https://stanforddaily.com)
- [privacy.claude.com](https://privacy.claude.com)
- [youtube.com](https://youtube.com)
- [biorxiv.org](https://biorxiv.org)
- [youtube.com](https://youtube.com)
- [youtube.com](https://youtube.com)
- [blog.adrq.dev](https://blog.adrq.dev)
- [huggingface.co](https://huggingface.co)
- [liquid.ai](https://liquid.ai)
- [aws.amazon.com](https://aws.amazon.com)
- [aws.amazon.com](https://aws.amazon.com)
- [openai.com](https://openai.com)
- [anthropic.com](https://anthropic.com)
- [thezvi.wordpress.com](https://thezvi.wordpress.com)
- [cnbc.com](https://cnbc.com)
- [youtube.com](https://youtube.com)
- [arxiv.org](https://arxiv.org)

## 2. Damra Vol 00:00:38

Zero is a strange number to get from two products people pay for. I'll hold the victory lap, though, because those are two lab harnesses running somebody else's model, which isn't what they were tuned around. What's the task?

## 3. Lenar Kess 00:00:51

It's a five-year value-added tax reconstruction across three generations of SAP: legacy ECC, a partial migration to S/4HANA, and a consultant's redesign. The agent gets raw system extracts. They have page headers and local number formats, the link tables are partial, and nothing documents how anything connects. The authors say it's what a team of specialists does over weeks. The agent has to reverse-engineer how three systems recorded the same business and reconcile to the cent. It also has to produce deliverables on company templates and hand over a reusable tool that gets tested on data it has never seen. Grading is deterministic, with no language model as judge, and it runs on Harbor.

## 4. Damra Vol 00:01:30

The no-judge part matters to me. A rubric graded by a model rewards a convincing story, and this task has an answer that exists in the data and can be checked to the cent. A harness can't charm its way to a score.

5. Lenar Kess 00:01:43

There are also traps built in that penalize invented findings. If a harness reports tax exposure that isn't there, it loses points for it.

6. Damra Vol 00:01:51

That's a good test for the failure you'd worry about most, an agent that gets tired and starts producing plausible numbers. Fine, what were the results?

7. Lenar Kess 00:02:00

OpenCode scored 0.542, and micro-cc, a small independent harness, scored 0.528. Claude Code and Codex got 0.00 each. The abstract says, quote, "No single step defeated any of them. What defeated them was inertia: keeping many weak signals alive across a long session." Then there's the frontier run. micro-cc and Claude Code on Claude Opus 5.5 scored 0.9986 and 1.000, in 22 and 25 minutes, for about five dollars eighty and five dollars eighty-nine.

8. Damra Vol 00:02:35

So with a frontier model the task collapses into twenty-five minutes and six dollars, and below the frontier the harness decides. I'd still push on the zero. Zero means no usable deliverable, so did they say what went wrong inside those two?

9. Lenar Kess 00:02:50

They trace the gap to concrete engineering: what goes back into the context, tools and defaults that assume the home model, recovery from a broken turn, and verification before declaring done. That's the level of detail I have. I'd add limits of my own. It's one task family in a first-version preprint, and the abstract gives one score per harness, so I'd want to see the spread across runs before I believed that 0.542 versus 0.528 means anything.

10. Damra Vol 00:03:17

The gap between half marks and nothing doesn't depend on that spread, though. It also matches a paper they cite, called Stop Comparing Large Language Model Agents Without Disclosing the Harness, which argues that on long tasks the harness often moves the result more than the model does. kw-bench is that argument with the model held fixed.

11. Lenar Kess 00:03:37

The cost line is the one I keep rereading. Quote: “In our runs the weeks of evidence work behind one tax decision cost between a quarter of a dollar and six dollars per harness-model pair. Dollars and cents, not millions.” And they say the specialist's job moves from doing the evidence work to holding the checklist and steering.

12. Damra Vol 00:03:55

There's a sentence earlier in the paper I like better: “You can outsource the making; you cannot outsource the understanding.” They're arguing that people automate the slide deck when they should be pointing the model at the ledger tracing underneath it.

13. Lenar Kess 00:04:08

Arize, the observability company we mentioned on Sunday for trace-based evaluation, put out two talks this week that read like the engineering behind that gap between zero and half. The first is on what they call Code Mode. Instead of loading a catalog of tool definitions into the agent's context, the agent searches for the tools it needs. It then runs them in a sandbox with Bash and Python, where Arize's REST endpoints and SQL are exposed.

14. Damra Vol 00:04:33

They report tokens down 43 percent, cost down 31 percent, and latency about flat. The gains hold across Claude Code, the Model Context Protocol, and command-line surfaces. They evaluated with Harbor, the same framework kw-bench uses, with binary verifiers. Those are Arize's numbers on Arize's tasks, so they're self-reported.

15. Lenar Kess 00:04:56

What makes the talk credible to me is that it lists its own costs. It names the sandbox lifecycle and network egress, then keeping secrets out of context and overbroad permissions. Debugging is on the list, along with what they call double sandboxing, durability, and weaker small models. That reads like a team that has actually run this.

16. Damra Vol 00:05:15

Secrets out of context connects to the rest of today. The moment the agent writes its own code against your endpoints, somebody has to decide whose credentials that code carries. We'll get there in the third segment.

17. Lenar Kess 00:05:27

The second talk is from Fuad, an AI product manager at Arize, about Alex, their in-product assistant. Alex 2.0 shipped in February on Sonnet 4.5. In June they merged the specialist agents

into a single agent with skills that load on demand.

18. Damra Vol 00:05:43

The to-do tool shows how much detail went in. It went through several versions, ending with the harness tracking the task list itself instead of trusting the model's own. And every turn ends with an explicit finish call. If the agent calls finish while items are still open, it gets bounced back, up to three times.

19. Lenar Kess 00:06:01

There's also a warning one step before the iteration limit, and then a forced finish. Compare that with the kw-bench list: recovery from a broken turn, verification before declaring done. It's the same territory, built by hand by a team that says it uses no agent framework.

20. Damra Vol 00:06:17

The memory design is my favorite. One markdown file per user, capped at 8,000 characters, rewritten by a background job, and they say it beat both vector retrieval and knowledge graphs. Large tool results get truncated with a pointer, and the agent queries them using a command-line JSON tool called jq.

21. Lenar Kess 00:06:35

Why would a capped file beat retrieval?

22. Damra Vol 00:06:38

My guess, and it's only a guess, is that a cap forces the rewrite to decide what matters, while retrieval keeps everything and ranks it at read time. They didn't say that. They just said it won.

23. Lenar Kess 00:06:50

Two arXiv papers this week point the same direction. Spine is a context runtime that organizes a long task as a tree of subtasks. SpineCodex, running on GPT-5.6 Sol, gets 11.8 percent more on SWE-Milestone at 32.4 percent lower cost. Against a recursive-subagent approach it scores comparably at 73.7 percent lower cost. The same idea in a harness called Pi gives gains between 9.5 and 17.5 percent, using Gemini 3.8 Flash and Qwen 3.8 at 27 billion parameters.

24. Damra Vol 00:07:25

The other is AgentBaton. It hands a task from a capable model to an efficient one partway through. Across 11 model pairs and two benchmarks it cuts API cost by 65.9 percent on average, with

accuracy preserved and in the best case up to 8.8 points higher. Going the other way, from efficient to capable, doesn't help.

25. Lenar Kess 00:07:46

That asymmetry is odd. I'd guess the early turns are where the approach gets decided, and a cheap model can follow a plan it couldn't have written. That's my read and the paper may say something different.

26. Damra Vol 00:07:57

Both are preprints, and I'd want someone else to reproduce them. But Spine's task tree and Alex's task list share an instinct, which is to take what the model would hold in its head and put it somewhere the harness can see.

27. Lenar Kess 00:08:09

Google DeepMind announced Genie 4, which it describes as a real-time interactive world model with native audio, built on Gemini. The Hacker News thread sits at 926 points and 534 comments, so it has drawn a lot of attention. Access is trusted testers for now, and the post says it's coming soon to Google AI Ultra subscribers in Project Genie.

28. Damra Vol 00:08:30

Native audio is the detail I'd poke at. If the sound is generated alongside the frames, then footsteps and water and engines have to follow what you do in real time, which would be harder than putting a soundtrack on a video. The post says it, but I haven't seen how it holds up outside the reel.

29. Lenar Kess 00:08:47

Here's what the post doesn't give. There's no resolution, no frame rate, no session length, and no latency figure. It makes a general claim about robotics, and I couldn't find a specific experiment attached to it.

30. Damra Vol 00:08:59

No latency number on something called real-time interactive is a big gap. Interactive could mean a game, or it could mean a slideshow that responds to key presses. And session length decides whether you can build anything on it, because I'd expect a generated world to drift the longer you stay in it.

31. Lenar Kess 00:09:15

That drift expectation is yours rather than theirs, and I'd keep it labeled that way. I agree the thread is reacting to the reel, so it tells us about the reel and not about the system.

32. Damra Vol 00:09:25

If the demos are representative, someone can type a place and walk into it with sound already running, and that's a different tool for a game prototyper or a filmmaker. Until there's a session length and a latency figure, I'm treating it as a very good demo.

33. Lenar Kess 00:09:40

John Feminella of Two Sigma opens a talk on agent controls with a banking story from 2020. In a Revlon loan payment, Citibank sent roughly 900 million dollars when it meant to send about eight million in interest. The cause was a set of checkboxes left unchecked in the payment interface, and every approver overrode the system's warning.

34. Damra Vol 00:09:59

That's a human in the loop approving everything because the warning looked like noise. It's useful as a caution for anyone planning to staff an agent review queue with tired people.

35. Lenar Kess 00:10:08

He lists five controls to prevent that. Three are in the transcript I have: an agent registry, review that scales with risk, and tool access that's granular and tied to a declared purpose. The transcript ends partway through the third, so I can't tell you what the other two are.

36. Damra Vol 00:10:23

Purpose-declared access is the interesting one. A service account says who you are, and nothing in it says why you're calling. An agent that's been prompt-injected can call for any reason it likes.

37. Lenar Kess 00:10:35

Keycard, a vendor in this area, gave a talk with the same diagnosis. Kamil argues that static tokens and service accounts can't attribute an agent's action to anyone or revoke it quickly. The fix they sell is a gateway. The credentials sit in a vault, and the gateway injects them per request after a policy check, so revoking an agent is instant.

38. Damra Vol 00:10:55

They sell it, so the talk is also a pitch. But it matches what Arize listed as a cost, keeping secrets out of context. If the agent never holds the credential, a leaked transcript doesn't leak the key.

39. Lenar Kess 00:11:06

Formal published a post on cloud agents querying production with OpenID Connect instead of long-lived secrets. It got 11 points on Hacker News. One commenter objects that OpenID Connect assumes the token's subject is the responsible party.

40. Damra Vol 00:11:21

That's the gap Feminella points at. The token proves which workload is calling, and says nothing about which person approved this action or what it was for.

41. Lenar Kess 00:11:29

Now the attack side. A paper called Log-To-Leak describes tool-description injection that makes the agent issue a follow-up call after the main task, and that call exfiltrates data. Attack success runs from 62.64 to 100 percent across 18 model-and-server pairs. On the GitHub server through the Model Context Protocol, with GPT-4o, attack success stays above 55 percent under three separate defenses.

42. Damra Vol 00:11:56

The conclusion is what I noticed. The authors say the host, the program running the agent, should authorize secondary tool calls. They aren't asking the model to resist. They're asking the layer around it to refuse.

43. Lenar Kess 00:12:09

One more, with an asterisk. Cross-User Poisoning, about agents that share context between users, showed up this week as a second version of a paper first posted a year ago, so it isn't new. Its finding is that task clustering and provenance prompting help, and that summarization alone doesn't.

44. Damra Vol 00:12:26

Summarization alone not helping should bother anyone shipping shared memory. A summary of poisoned context is still poisoned, just shorter.

45. Lenar Kess 00:12:34

Two Anthropic items today that sit on opposite sides of training. The Stanford Daily reports that professors' books were pirated to train Claude. The allegation comes from plaintiffs in a case called Cruz v. Anthropic, and the article reports no response from Anthropic. It follows the Bartz settlement, which was 1.5 billion dollars.

46. Damra Vol 00:12:54

So this is a plaintiff's account of what was done. I'd keep that in front, because the article is reporting a complaint and not a finding.

47. Lenar Kess 00:13:01

Right. The other is Anthropic's privacy help page, titled whether your data is used for model training, which reached 178 points on Hacker News. One commenter claims that dismissing the feedback prompt may count as giving feedback. I haven't verified that, and I'd treat it as one reader's interpretation of the page.

48. Damra Vol 00:13:20

A claim like that is what a page like this gets scrutinized for. If it's true, the default is doing more than the toggle.

49. Lenar Kess 00:13:27

Then there's the Latent Space interview with Anthropic, an hour and 46 minutes. It covers Opus 5.5, which responded to complaints that Opus 5's writing had become convoluted, and which also cut prices. It also covers a paper that extracted reasoning traces. Anthropic says it was responsibly disclosed and patched.

50. Damra Vol 00:13:45

What I'd quote from the interview is where they're heading. They're moving toward keeping thinking blocks across turns and locking thinking to a prefix, and they want limits on cross-account abuse. They also shipped compaction on the API side, so harness maintainers can compress a long session without dropping the model's own reasoning.

51. Lenar Kess 00:14:04

Their stated reason is primarily safety. A distilled model runs without Anthropic's output classifiers, so copying reasoning traces produces a capable model with none of the checks. They describe the commercial impact as secondary.

52. Damra Vol 00:14:16

That's the stated reason, and safety and commercial protection point the same direction here. I can't separate them from outside, and neither can you.

53. Lenar Kess 00:14:24

There was also an architecture note I liked. They described a brain-and-hands design, where the sandbox can die without killing the agent. That's a direct answer to the sandbox lifecycle cost Arize

listed.

54. Damra Vol 00:14:36

If your agent's state lives outside the box it's running in, a dead sandbox is a retry and not a lost afternoon. For anyone building a harness, that plus API-side compaction is the practical news.

55. Lenar Kess 00:14:48

A bioRxiv preprint from a group connected to Galaxy, the bioinformatics platform, ran 30 unguided local agents on the same dataset. They produced 18 incompatible analysis pipelines. The abstract also reports reproducing published studies on public infrastructure. It says that exposed unstated modeling choices, reference mismatches, and tutorial inputs that had been subsampled.

56. Damra Vol 00:15:12

My read is that eighteen pipelines from thirty runs means the agents disagree about choices a human analyst makes without writing them down. Which normalization, which reference genome. The agents surface those choices because each run has to make one.

57. Lenar Kess 00:15:27

That's plausible, though the abstract doesn't say how many of the 18 differ in ways that change the biological conclusion, and I'd want that number.

58. Damra Vol 00:15:35

There's a 51-second Machine Learning Street Talk clip on a related problem in mathematics. The speaker says autoformalization floods Mathlib with imprecise Lean, meaning statements that compile but don't say what the mathematician meant.

59. Lenar Kess 00:15:49

It's a clip, so I'd treat it as a speaker's assertion. But it matches the bioRxiv result: the output passes the checker and is still wrong about intent.

60. Damra Vol 00:15:58

That's kw-bench's argument from the other side. They built a task where the answer is checkable to the cent so that accepted-but-wrong can't score.

61. Lenar Kess 00:16:07

A few smaller things. Nate B Jones has a video with four principles for working with agents. Count the whole human cost of reading the output. Run more experiments than you commit to. Set an end

date before making an agent recurring. And give a chief-of-staff agent an attention budget.

62. Damra Vol 00:16:23

The end-date example is the best one. He describes a weekly market update that kept arriving two months after he'd moved on. He also describes a chef, Dan Ninarol, who he says made about 20,000 dollars in around six weeks, and a painter, Tony Severino, who caught an agent using a floor-area multiplier instead of wall measurements.

63. Lenar Kess 00:16:44

There's a Hacker News thread of 148 points on a blog post called Here be monsters, about what you still need to know when agents do the work. One commenter there says database schemas are what language models get subtly wrong. I'm attributing that to the thread, since I couldn't confirm the blog itself.

64. Damra Vol 00:17:01

Liquid AI released LFM3, a text model with 1.7 billion parameters on what they call a tri-hybrid architecture, aimed at local agents. There are no independent benchmarks yet, so for now it's a claim plus an artifact on Hugging Face.

65. Lenar Kess 00:17:18

Two AWS case studies. Exowatt reports up to 49 times lower cost, and that's the customer's number. Property Finder's is the more interesting one: 96.8 percent of its evaluation labels were inherited from the incumbent model, and it cut cost 86 percent.

66. Damra Vol 00:17:33

That 96.8 is the detail. If the labels come from the model you're replacing, the test measures agreement with the incumbent, which isn't the same as being right. AWS's post is about fixing evaluation bias, so they know.

67. Lenar Kess 00:17:48

On distribution, OpenAI announced Oracle apps in ChatGPT, Anthropic announced an expanded partnership with the General Services Administration, and OpenAI published a customer story on Legora. A Hacker News commenter on the GSA item asks about the Defense Department's supply-chain-risk designation. That's unverified, so I'm only mentioning that the question was asked.

68. Damra Vol 00:18:09

Zvi wrote up Gemini 4 Argon and its safety cards. His summary, as passed along in the newsletter, is rising evaluation awareness, and that it can bypass chain-of-thought monitoring, though doing so hurts performance on complex tasks. That's Zvi's reading, and I haven't checked the cards.

69. Lenar Kess 00:18:26

CNBC reports Meta is doing a 95 billion dollar stock offering tied to its AI spending. I only have the headline-level detail, so that's all I'll say.

70. Damra Vol 00:18:35

Last one is from Chainguard. Patrick Smith cites Mandiant data showing time-to-exploit going from a 63-day buffer in 2018 and 2019 to negative 70 days by 2024, meaning exploits now arrive before patches. That's the speaker's claim, and Chainguard sells the products he recommends, but the direction matches what the credential papers describe.

## Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic