

Gartner: Only 13% of CIOs See Big AI Value, and Apollo Finds No Payoff Outside Tech

2026-10-25 / 00:21:13

“The AI story will be decided in the S&P 493, not the Mag 7.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

Gartner says 13 percent of CIOs have seen significant AI value, and Apollo's chief economist says there's no payoff outside tech, so the episode starts with whether the spending has a customer. After that: agents as attackers, why code review is the new bottleneck, three talks on trusting what agents produce, a volunteer cleanup of OpenAI's math write-ups on Terence Tao's blog, and shorter items on Qwen, Cerebras and a cytometry preprint.

- [Apollo: no signs of AI payoff outside the tech sector](#)
- [Gartner CIO survey press release](#)
- [NLW's AI Daily Brief on neoclouds and capex](#)
- [Steal My Bitcoin](#)
- [Dan Luu: We should be more ambitious with the software we build](#)
- [Simon Willison on building a strategy game with agents](#)
- [OpenAI Math Revised, a community hub on Terence Tao's blog](#)
- [Observer: Terence Tao on AI and mathematics](#)
- [Cerebras: deploying trillion-parameter models](#)

- [GVS5H](#)
 - [bioRxiv: language models annotating mass cytometry clusters](#)
-

SEGMENTS

- [00:00:04](#) Is the AI payoff showing up?
- [00:05:22](#) Agents as attackers
- [00:08:33](#) Code is cheap, review isn't
- [00:11:19](#) Trusting what agents produce
- [00:14:33](#) A cleanup of OpenAI's math write-ups
- [00:16:44](#) Shorter items: Qwen, Cerebras, cytometry

Transcript

1. Lenar Kess 00:00:04

Gartner put one question to CIOs this month: has your enterprise realized significant or transformational value from AI? Thirteen percent said yes. That's from a press release dated October 19th, and this weekend an Apollo note followed with a title that doesn't leave much room: no signs of AI payoff outside the tech sector. So if you're financing data centers on the assumption that every company eventually buys AI, how long can you wait for the other eighty-seven percent?

- [youtube.com](#)
- [apollo.com](#)
- [gartner.com](#)
- [youtube.com](#)
- [stealmybitcoin.com](#)
- [youtube.com](#)
- [danluu.com](#)
- [simonwillison.net](#)
- [v.redd.it](#)
- [v.redd.it](#)
- [youtube.com](#)

- youtube.com
- youtube.com
- terrytao.wordpress.com
- observer.co.uk
- i.redd.it
- github.com
- v.redd.it
- reddit.com
- cerebras.ai
- youtube.com
- youtube.com
- biorxiv.org
- i.redd.it
- x.com
- twitter.com

2. Damra Vol 00:00:30

And the odd part is that the people who track the supply side aren't panicking. I watched NLW's Daily Brief episode on this, and his argument is that the stress indicators people keep pointing at mostly have physical explanations: concrete, power hookups, construction schedules.

3. Lenar Kess 00:00:46

That's where we start. After it: agents as attackers, with a PHP bug, a certificate issuer that exposed an absurdly weak private key, and a bitcoin bounty nobody has claimed. Then a talk on why code review is now the bottleneck, three practitioner talks on trusting what agents produce, Terence Tao's blog hosting a cleanup of OpenAI's math write-ups, and a round of shorter items including Cerebras and a cytometry preprint.

4. Damra Vol 00:01:12

Start with Apollo, then, because it's the most concrete.

5. Lenar Kess 00:01:15

Torsten Slok, Apollo's chief economist, wrote on Saturday that, and I'm quoting, 'the technology has delivered zero financial payoff to corporate America outside of tech.' His charts show revenue per worker and profit margins climbing for the Magnificent Seven and sitting flat for the other 493 companies in the S&P 500. He also cites Ramp data showing that token spending is concentrated in a small number of firms.

6. Damra Vol 00:01:40

What caught me is the last stretch of his argument. He says mainstream companies are starting to treat token spending as a cost center that cuts margin, not something that grows revenue, and that they don't need the newest models and won't pay for them. That's aimed straight at the premise behind frontier pricing. If the S&P 493 buy the cheap tier, the revenue forecast for the expensive tier has a problem.

7. Lenar Kess 00:02:03

He gets a good line out of it, too: 'the AI story will be decided in the S&P 493, not the Mag 7.' And his reason is that 'in business, intelligence is mostly judgment.' I'd separate the instruments, though. Gartner's number is CIOs describing their own companies, and Slok's is financial statements. A CIO saying the value hasn't been transformational isn't the same as margins staying flat.

8. Damra Vol 00:02:28

Right, and the third instrument is the one NLW cares about, which is whether the people lending money for GPUs get paid back. He walks through a Gartner forecast of roughly 8.5 trillion dollars of revenue needed by 2030 to cover announced capital spending and financing costs, against about 3.6 trillion in forecast enterprise and consumer AI revenue. Out of that gap comes a claim that eighty percent of neocloud providers fail. Those are forecasts, and the eighty percent is a claim riding on them, so I'd hold it loosely.

9. Lenar Kess 00:03:01

His counter is Blackstone. Their executives argue neocloud debt is safe because the tenants are hyperscalers, which are creditworthy, and the capacity is being built against demand that already exists. NLW sorts what could go wrong into three scenarios. One is a neocloud default where no creditworthy tenant stands behind the loan. Another is a deterioration in a tenant's own credit. The third is hyperscalers cancelling leases because demand is weak, as opposed to because power showed up late.

10. Damra Vol 00:03:29

That last distinction matters because a cancellation looks identical in a headline either way. Then there's the Amazon story, undeployed Nvidia GB200 racks sitting in inventory, which people read as channel stuffing. NLW reads it as lag. Data center construction and grid connections are running behind. Staging GPUs ahead of the racks is how you hold an allocation while pricing premiums are still high. I buy that. Idle racks waiting for power are a construction schedule.

11. Lenar Kess 00:03:59

He also puts a size on the contagion worry. About 740 million dollars of CoreWeave debt is held by large insurers and pension funds, which he treats as a negligible slice of portfolios that big, and nothing like 2008. We went through the draft Anthropic prospectus earlier this week, and the number I keep wanting next to it is utilization, because that's what tells you whether the spending is a response to demand.

12. Damra Vol 00:04:22

So which one shows first? My guess is none of NLW's three, at least not at the start. Slok's version is slower. The buyers exist but the margins don't, so the pressure shows up as customers renewing at lower tiers. That's harder to see than a default, and lenders won't care until it reaches a tenant's credit.

13. Lenar Kess 00:04:40

There's a counterpoint inside the same episode. A KPMG study of more than 500 early-career professionals found that AI output quality tracked how people worked with the model, with the high performers iterating, evaluating and refining, more than it tracked what they knew about the domain. If that holds up, a thirteen percent figure might be measuring management practice as much as model capability.

14. Damra Vol 00:05:02

Which is the optimistic reading of Gartner's number, and I like it, because it explains Slok's flat margins without needing AI to be useless. A tool that rewards a particular workflow shows up first in the companies that happen to have that workflow. I'd still rather bet on the workflow gap closing than on Slok's view that judgment stays out of reach.

15. Lenar Kess 00:05:22

Nate B Jones spent twenty-five minutes on this week's security incidents, and I'd start with LightLLM. By his account, an October release of it contained credential-stealing malware while the tagged source code in the repository was clean. The malware came in through a build job, and that job still had access to production keys. An investigator reading the repo sees nothing wrong.

16. Damra Vol 00:05:43

Which makes the usual check, diff the release against source, useless, because the evidence is in what actually executed. Nate's advice follows from that: treat any compromised service as a whole-environment breach, rotate every secret, and audit execution history instead of trusting the repo. He also suggests testing any model you'd use for an investigation on a known incident, to see whether it stops at the clean source or digs into build jobs and deployed artifacts.

17. Lenar Kess 00:06:11

The other two incidents, as he tells them. PHP shipped fixes for critical bugs that allow code execution through PHP's Apache module before application logic even runs, and UK and EU agencies recommended isolating affected systems right away. And a certificate issuer in India exposed eight trivially weak private keys, one of them literally the number four. The researcher Scott Helme reportedly recovered it in under a second using Claude.

18. Damra Vol 00:06:38

A private key that's the number four is funny for about two seconds, and then you remember that certificates chain. One weak key at an issuer is a problem for everyone who trusts that issuer, which is Nate's point about cascading trust. What strikes me is the speed. The cost of looking dropped to nearly zero, and nobody had looked at that key.

19. Lenar Kess 00:06:58

That's his thesis: patching on a maintenance cycle is too slow when agents keep finding latent bugs. He cites Anthropic as reporting over 29,000 candidate flaws across six months of scanning, with roughly 6,000 triaged by hand, and that's his retelling of the figures, so I'd treat them as his citation. His prescription is to use the strongest model you can for defense, and he argues that model refusals often get in the way of emergency response more than they stop misuse.

20. Damra Vol 00:07:25

I'd argue with that. 'Strongest model for defense' is easy to say when you're the defender with a budget, and the attacker gets the same model. The asymmetry he describes, where an attack is cheap and one person can scale it, cuts both ways. That's why I like the Hacker News item as a counterweight.

21. Lenar Kess 00:07:42

'Steal My Bitcoin' is a Show HN that reached 486 points and over 300 comments. As the title says, it's a standing public bounty. The top comment treats the fact that it's still unclaimed as evidence against the story that swarms of autonomous agents are breaking everything.

22. Damra Vol 00:07:59

It's a decent test because it's public and falsifiable, which most claims about what agents can do aren't. But one unclaimed target isn't a survey. Whether an agent can break one well-defended system says little about the thousands of unmaintained ones with keys that trivial. The attack surface Nate worries about is the long tail, and the bounty is a single target at the strong end.

23. Lenar Kess 00:08:21

I take both. The bounty says the best-defended targets are still hard, and that certificate key says the worst-defended ones were never defended. How many issuers and build pipelines look like that is a number nobody has.

24. Lenar Kess 00:08:33

Laurie Voss, who co-founded npm, gave an AI Engineer talk built on one statistic: developers using autonomous coding agents produce 741 percent more code and ship about 30 percent faster. Those are his citations, not mine. His explanation is review. Humans read code at about 450 lines an hour, which makes reading an agent's diffs a losing proposition.

25. Damra Vol 00:08:56

And he doesn't let automated reviewers off. Multiple AI reviewers flag different bugs with no consensus, and he cites mergeability benchmarks where pass rates fall from 87 percent to 54 percent when you judge against something stricter than the standard test suite. So the tool meant to relieve the bottleneck has a variance problem of its own.

26. Lenar Kess 00:09:15

Then he applies Goldratt's Theory of Constraints. Fix review and the constraint moves downstream to deployment, scaling and security operations. Automate those too, and assume demand for software is effectively unlimited, and you end up at defining the problem: what 'good' means in a given domain. He pairs that with Fred Brooks's accidental versus essential complexity, and with the idea that code is a lossy projection of a specification.

27. Damra Vol 00:09:40

I stop agreeing at 'permanent bottleneck.' His own evidence cuts against it. He cites Cursor's BugBot turning feedback into 44,000 automated rules, which is somebody's tacit knowledge becoming explicit. Anything you can write down or cheaply verify gets absorbed, and rules extracted from feedback are exactly that: writing it down.

28. Lenar Kess 00:10:00

He'd answer that the unstated business need is by definition what hasn't been written down, knowledge that lives in people and not in training data. Dan Luu's essay, 'We should be more ambitious with the software we build,' reached 479 points on Hacker News, and its top comment goes at the same nerve from the other side. It complains about people reporting that AI replaced what they used to do, instead of what new things they built.

29. Damra Vol 00:10:25

That's the sharper version of Voss's argument for me. If the constraint is knowing what's worth building, then 741 percent more code aimed at the same backlog is a symptom. Simon Willison's write-up goes the other direction: he built a mobile real-time strategy game over 321 agent commits and wrote up what he learned. A game is software nobody had a ticket for.

30. Lenar Kess 00:10:47

The Claude subreddit has more of those this weekend. One person is twenty days into building a North Atlantic island in Unity with Claude Code, and another built an open-source laser dry-fire shooting game where a webcam watches your monitor and scores the hits. None of them started as anyone's roadmap item.

31. Damra Vol 00:11:04

Which makes me think Luu is right that the ambition is available and the review cost is what's in the way. A hobbyist can accept a game that's sometimes wrong. Voss's 450 lines an hour only bites where a wrong merge costs money.

32. Lenar Kess 00:11:19

Three AI Engineer talks in this batch arrive at the same problem from different directions, and the first has a clear example. A forward-deployed engineer at Arize described a one-line prompt change on a support agent that improved action safety by 17 points and reduced task completion by about the same amount. Net result: zero. Nothing in the code diff shows it.

33. Damra Vol 00:11:39

That's the whole case for why a diff is the wrong instrument for agents. A prompt edit is a tiny diff with an enormous behavioral footprint, and the speaker's position is that production traces are the source of truth. The loop has five steps. You build a dataset from real traces, including known failures, and then build scorers, which can be a model as judge, deterministic checks, or trained classifiers. After that you run a baseline, iterate offline, and finally gate pull requests in continuous integration on the scores.

34. Lenar Kess 00:12:10

Arize sells the platform in that demo, naturally. The loop works with any tooling, though, and the trade-off number is theirs: you can improve one dimension and silently lose another. That leads into the benchmark talk, which bothers me more.

35. Damra Vol 00:12:24

Google DeepMind and Kaggle are turning Kaggle into an evaluation platform for frontier models, and the product manager leading it said they reproduced 65 industry benchmarks and found problems. They found score manipulation through the container the agent ran in, and evaluation and grading agents that ran together instead of being kept separate. They also saw run-to-run variance of up to plus or minus twenty percent on identical runs, and mismatched default reasoning effort between models.

36. Lenar Kess 00:12:51

Plus or minus twenty on identical runs means a lot of leaderboard gaps sit inside the noise. Kaggle's pitch is sandboxed execution, logged proxy calls, open tracing and containerized grading, with Epoch named as an independent auditor, free compute and 5 million dollars in grants this quarter. It's also a promotion of Kaggle's own infrastructure.

37. Damra Vol 00:13:11

Which is fine, and the finding doesn't depend on the pitch. If you've quoted a benchmark number to anyone this year, ask how many runs it was and what reasoning effort was set. They also previewed MandateBench, an evaluation for agents doing buy-side equity analyst work. It grades whole outputs with rubrics, from the fund mandate through to the investment memo. They showed results on 10 of 50 tasks, with the full suite due in November, so I won't rank anything from it.

38. Lenar Kess 00:13:39

The third talk is Kepler, whose CEO described a design for financial research where the language model plans queries and never writes a number. Code does the arithmetic, every figure carries a link to its source document, and anything uncited gets stripped before it reaches the user. By his account they draw on over 30 million SEC documents.

39. Damra Vol 00:13:59

I like how little it asks of the model. Instead of making the model trustworthy, you make its trustworthiness irrelevant for the numbers. A claim without a source can't be shown. He states the price plainly: it suits mission-critical financial analysis and not marketing copy or speculation, because the stripping filter removes exactly what makes speculation speculation.

40. Lenar Kess 00:14:20

Put next to Voss, that's a practical answer to what the human bottleneck is. The human defines what good means, and a system like Kepler's or Arize's encodes it as a check that runs every time. Writing that check well is still the human job.

41. Lenar Kess 00:14:33

Terence Tao's blog posted something on Saturday that's easy to misread from the Hacker News headline. It's a guest post, not something Tao wrote himself. The authors are Abhinav Nellore and Johannes Schmitt, along with Soledad Villar and Rachel Ward. They're announcing a community project called OpenAI Math Revised, which aims to improve the mathematical write-ups OpenAI published in its own GitHub repository. Their reason, in their words: those documents 'can be difficult to understand and can have incomplete citations.'

42. Damra Vol 00:15:04

So the claim isn't that the math is wrong. It's that OpenAI got results and published them in a form the mathematical community can't easily check or cite, and so students, postdocs, professors and enthusiasts are invited to rewrite them into something readable. The comments are skeptical on exactly the labor question: why are unpaid volunteers polishing a for-profit company's output, and shouldn't they write their own papers?

43. Lenar Kess 00:15:29

That objection has a point and also misses one. The authors presumably want the results usable as mathematics, and that requires references. A proof you can't trace to prior work isn't something other people can build on, whoever wrote it.

44. Damra Vol 00:15:43

Which is the same demand as Kepler's, every claim pointing at a source. In mathematics it's called a citation, and it's the whole trust model of the field. Then there's the Observer piece, headlined 'Terence Tao: AI turns beautiful mathematics into toxic waste.' Going by the headline and sub-headline, it's about Silicon Valley rushing at long-standing problems in a way that leaves out elegant solutions and treats mathematics as training material.

45. Lenar Kess 00:16:09

I'd keep those two apart. The hub is practical, and the guest authors run it, not Tao. The Observer headline is a stated value judgment, a view about what's lost when the goal is an answer instead of understanding. People can disagree with that without disputing a single theorem.

46. Damra Vol 00:16:24

And a screenshot circulating on the singularity subreddit claims labs are working to replace research economists with agents that produce publishable answers in under an hour. It's an unverified message, so I'd leave it as a rumor. It does point at the same pressure in a different field: output arrives faster than the community's ability to check it.

47. Lenar Kess 00:16:44

Quick run through the rest. A GitHub project called GVS5H claims five Qwen3.8 27-billion-parameter models, run through its harness, match Fable 5 on a coding benchmark. It got 146 points on Hacker News, and the top reply points out that the benchmark is saturated, dates from early 2025, and has no comparison against current frontier models.

48. Damra Vol 00:17:06

That reply cuts the claim down to size. A result like that is easier to hit on a saturated test, where the ceiling is close, and I only know the method as far as the repo shows it. I'd want it rerun on a harder benchmark.

49. Lenar Kess 00:17:19

Two hardware tricks showed up on the LocalLLaMA subreddit. Lean-QwenFlash reports a 125-billion-parameter Qwen3.8 Flash model at about 20 tokens a second on 16 gigabytes of video memory and 8 gigabytes of system memory, by streaming from an NVMe drive. And another post says Qwen3.8 27B runs at a 192 thousand token context window on the Neural Engine of an M6 Mac Mini with 32 gigabytes. Both are self-reported.

50. Damra Vol 00:17:48

The SSD one is the one I'd play with. Twenty tokens a second from a drive means the bottleneck moved from memory capacity to read bandwidth, and a mixture-of-experts model only touches a few experts per token, which is why the trick can work at all.

51. Lenar Kess 00:18:02

Cerebras published how it serves trillion-parameter models, and the post reached 204 points. The constraint is memory. One of its wafers carries 44 gigabytes on-chip, and a model like the roughly three-trillion-parameter Kimi K3 at 4-bit weights needs about 1.5 terabytes, by the post's numbers and my arithmetic. Their answer is to keep whole layers on a wafer so the weights never move: 59 wafers, with only activations passing between them over Ethernet at about 2 microseconds a hop.

52. Damra Vol 00:18:32

The layout detail I like is that the 24 dense-attention layers get a wafer each, so the leftover memory holds the key-value cache, while linear-attention layers need only a small fixed state per request, so two share a wafer. That means concurrent users are capped by cache space on those 24 wafers. It's a vendor post, aiming at 1,000 to 2,000-plus tokens per second, with no independent throughput numbers.

53. Lenar Kess 00:18:57

Bridgewater and Browserbase gave AI Engineer talks on data access and retrieval for finance agents. Bridgewater's claim is that human readers and agent readers need different retrieval tuning: humans read about 300 words a minute and need precision, agents read about 30,000 and want recall. They took precision at 10 from 60 percent to between 90 and 95 percent on about 200 hand-curated examples. Browserbase offers a forked Chromium for login-walled portals, plus Stagehand cutting a page from around 122 thousand tokens to about 4 thousand as an accessibility tree. The 30,000 saved hours at Ramp is Browserbase's own figure.

54. Damra Vol 00:19:35

The 300 versus 30,000 is a good number to design around, even as a round figure. It says an agent can afford to read a lot of junk you'd never show a person.

55. Lenar Kess 00:19:45

A bioRxiv preprint has language models label mass cytometry clusters from marker profiles in a single call. Claude Sonnet 4.6 matches the reference annotation at the major-lineage level 85 to 98 percent of the time, and the best local model gets 83 to 96. Claude Opus 5 was no more reliable than Sonnet. It's a preprint, four datasets, one lab.

56. Damra Vol 00:20:08

They had two experts annotate blind. The experts agreed on about 99 percent of major lineages but only about 73 percent of T-cell maturation states, and on that axis the best models fall inside the expert disagreement range. So 'matches the reference' depends on how much the experts agree with each other.

57. Lenar Kess 00:20:27

One more, at the size it is. Users on the Claude subreddit report a refusal when asking for help reverse engineering a smart TV, one that didn't fire for a video game. Another post claims the same happened to a Reverse Engineer Anything project a day after launch. Neither Anthropic nor OpenAI has said anything in these sources, and a separate post about a Samsung TV running a TensorFlow Lite model on a 960 by 540 image has a thin discussion.

58. Damra Vol 00:20:53

It's screenshots, so it could be a classifier side effect and not a policy. It does echo Nate's complaint from earlier, that refusals get in the way of legitimate security work. Until Anthropic says whether anything changed, the claim is one refusal on a TV and none on a game.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic