The Million-Token Bill Arrives

2026-06-20 / 00:23:04

> *“A million-token window only becomes useful memory if the bill, the cache, and the feedback loop all survive contact with the work.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

DeepSeek previewed a V4 model family with one-million-token context claims on the same morning that corporate AI spending discipline, coding-agent evaluation papers, and robotics ownership news pointed at a shared practical constraint: intelligence is getting measured by the loop around it.

- DeepSeek-V4 preview paper claims one million tokens for both Pro and Flash, with lower long-context inference FLOPs and key-value cache use than DeepSeek-V3.2; the open question is how much retrieval and agent memory this displaces in practice.
- Financial Times on companies reining in AI usage supplies the budget counterweight: once finance teams meter model usage, pilots become governed systems.
- The grite paper, StaminaBench, Predictive Validity for LLM agents, and AutoPass turn agent evaluation away from single-task demos and toward coordination, stamina, out-of-sample rank stability, and runtime evidence.
- Startup Fortune on Hyundai and Boston Dynamics reports Hyundai buying SoftBank's remaining 9.65 percent stake for $325 million, which makes the robotics

question less about demos than factory ownership, reliability, and patient deployment.

- ENPIRE and Human Universal Grasping give the technical robotics companion: real-world policy self-improvement, reusable reset and verification loops, and human grasp data at million-frame scale.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   DeepSeek posted a V4 preview paper today, Saturday, June twentieth. The family has two mixture of experts models: DeepSeek-V4-Pro, at 1.6 trillion total parameters with 49 billion active, and DeepSeek-V4-Flash, at 284 billion total parameters with 13 billion active. Both are described as supporting a one million token context window. That is the fact to start with, because everything else in the paper depends on whether the long window is affordable enough to use.

   - arxiv.org
   - ft.com
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - startupfortune.com

- arxiv.org
- arxiv.org
- x.com
- reddit.com
- reddit.com

2. Damra Vol 00:00:39

   And the paper doesn't just say one million tokens in the marketing sense. It gives a mechanism. DeepSeek says V4 combines compressed sparse attention with heavily compressed attention, then adds manifold-constrained hyper-connections and the Muon optimizer. The claim isn't just, we made the window bigger. The claim is, we changed the cost curve enough that the window can be routine.

3. Lenar Kess 00:01:05

   Right. In the one million token setting, DeepSeek says V4-Pro uses 27 percent of the single-token inference FLOPs and 10 percent of the key-value cache compared with DeepSeek-V3.2. Flash is even more aggressive: 10 percent of the single-token FLOPs and 7 percent of the cache against V3.2. I would keep those as claims from the paper, not independent truth yet, but they are specific enough to argue with.

4. Damra Vol 00:01:33

   The key-value cache number changes how I read this. A long context window that eats the machine alive becomes a demo feature. A long context window with a smaller cache becomes a different memory primitive. You can imagine an agent carrying a codebase, a run history, review notes, logs, and several design documents without immediately turning every task into retrieval triage.

5. Lenar Kess 00:01:58

   Although I would still be precise about the word memory there. A million tokens doesn't mean the model understands everything in the window, or that it will keep old instructions and current requirements in the same priority order. The paper is strongest when it talks about architecture and efficiency. It is less settled when it implies that longer context by itself makes long-horizon work reliable.

6. Damra Vol 00:02:21

   That distinction matters for builders. If you use retrieval today, you're usually deciding what evidence to bring into a smaller window. If the window expands, you still need a policy for what the model should attend to, what it should ignore, and when it should reread the source file instead of

trusting its compressed history. The attention bill went down in DeepSeek's telling. The judgment problem did not disappear.

7. Lenar Kess 00:02:46

   The paper also says the models were pre-trained on more than 32 trillion tokens and then post-trained through specialist training and on-policy distillation. So the model story isn't only the attention architecture. DeepSeek is trying to make V4-Pro the stronger general model and V4-Flash the cheaper high-reasoning option, especially when the prompt gets very long.

8. Damra Vol 00:03:11

   It is still a preview. That word should stay visible even if we don't say it with a raised eyebrow. The checkpoints are available, the paper points at benchmark results, and the architecture details are concrete. But when a paper says a model is three to six months behind frontier closed models on some reasoning tasks and ahead on some long-context tasks, I hear an engineering claim, not a coronation.

9. Lenar Kess 00:03:37

   My read is pretty simple: this is an important open-model artifact because it gives long context a plausible engineering path. It doesn't end retrieval, and it doesn't make agent state easy. It makes a new bet more attractive: maybe the next agent architecture spends less energy deciding which tiny slice of the past to remember and more energy proving that its current action matches the actual state of the work.

10. Damra Vol 00:04:01

    The test I would give it is practical: give it a messy repo, a multi-day agent session, a bunch of logs, and a rollback request. Then ask whether the model finds the one constraint that still governs the task. Long context is exciting when it improves that answer. Otherwise it is a bigger attic.

11. Lenar Kess 00:04:21

    The Financial Times has a story today under the headline, quote, 'We created a monster': companies rein in AI usage as costs strain budgets. The full article is paywalled from my side, so I'll stay with the visible headline and the source summary I have. Amazon, Walmart, Cisco, Uber, and other companies are reportedly tightening or pulling back AI usage as deployment costs hit budgets.

12. Damra Vol 00:04:47

    That is the necessary counterweight to the DeepSeek item. Every long-context release sounds like relief until someone asks how many employees can paste a million tokens into a reasoning model

every day. A finance team doesn't care that the context window is elegant if the usage graph turns into a surprise line item.

13. Lenar Kess 00:05:07

    And this isn't an argument that AI doesn't work. It is closer to the moment when a tool leaves the pilot budget and enters the operating budget. In the pilot phase, usage often means curiosity: try the expensive model, run the agent again, let the assistant chew through the whole workspace. In the operating phase, usage means approval paths, caps, internal gateways, and someone asking whether the model choice matched the task.

14. Damra Vol 00:05:34

    The model router becomes a political object inside the company. Who gets the strongest model by default? Which workflows get long context? Which teams have to justify a reasoning run? And who owns the uncomfortable answer when the cheap model makes a mistake that the expensive model might have avoided?

15. Lenar Kess 00:05:53

    That last question is why I don't like the simple version of the cost story. The simple version says companies overspent on AI and now the grown-ups are saying no. Sometimes, sure. But a better version says companies are discovering that model intelligence has to be allocated. You don't want the executive assistant summarizing a lunch menu on the same budget path as an agent modifying payment code.

16. Damra Vol 00:06:18

    You also don't want a policy that says everyone gets the cheap model because the bill was embarrassing. That creates a shadow system. People paste sensitive work into consumer tools, or they keep rerunning low-quality outputs until the total spend is worse. Cost control that ignores work quality just moves the cost into review, rework, and risk.

17. Lenar Kess 00:06:41

    The DeepSeek efficiency claim and the FT cost story belong near each other, but lightly. If long-context inference gets cheaper, some usage restrictions loosen. If companies learn to meter usage intelligently, some expensive models become easier to justify. The craft problem is to make the meter legible enough that a builder can choose the stronger tool when the work demands it without turning every request into a procurement hearing.

18. Damra Vol 00:07:09

And that means the logs matter. Not surveillance theater, just enough record to answer: this task used this model, with this context size, for this outcome. If the organization can't connect spend to results, finance will govern by invoice shock. If it can, the conversation moves to thresholds, not panic.

19. Lenar Kess 00:07:29

    Allocation is the word I keep coming back to. The next round of AI tooling isn't just smarter models. It is smarter allocation of context, reasoning, retries, and review. The bill arriving is annoying, but it is also the moment the tool has to become part of the system instead of an exception to it.

20. Lenar Kess 00:07:49

    A cluster of agent papers today is more interesting as a group than as four separate headlines. One paper introduces grite, a git-native coordination substrate for concurrent coding agents. Another introduces StaminaBench for long multi-turn coding sessions. A third argues that agent benchmarks should be judged by predictive validity, meaning whether rankings transfer out of sample. AutoPass applies multi-agent language-model workflows to compiler tuning with compiler and runtime evidence in the loop.

21. Damra Vol 00:08:21

    That is a good bundle because none of those papers is asking, can the model solve one impressive task? They are asking whether the surrounding process survives. Do agents collide over work? Do they degrade after turn five? Does a leaderboard predict the deployment case? Does the compiler tuning loop verify the speedup on actual hardware?

22. Lenar Kess 00:08:42

    The grite paper starts from a nice empirical discomfort. Autonomous coding agents are producing pull requests faster, but they are accepted less often in large-scale studies. The authors say pull request history misses what happened before the pull request: duplicate work, conflicting edits, abandoned attempts, and races to close the same task. So grite stores coordination records inside git as an append-only, optionally signed event log with conflict-free replicated data type semantics and advisory leases.

23. Damra Vol 00:09:14

    Their log is both the product and the measurement instrument. In the synthetic agent sweep, duplicate work falls from 78 percent to zero when agents use leases plus shared state, while useful throughput more than triples. I wouldn't generalize that magnitude to real coding agents yet; they are explicit that the main run uses deterministic seeded agents. But the category is right. The work that never becomes a pull request can still waste the day.

24. Lenar Kess 00:09:43

    StaminaBench is the other side of that. It asks how many consecutive change requests a coding agent can handle before failing. The setup is a generated REST API server that evolves through 100 follow-up changes, with programmatically generated tests and black-box HTTP evaluation. Their sharp result is that without a useful feedback loop, all tested models fail within about five or six turns. With detailed test feedback and retries, performance improves by as much as 12 times.

25. Damra Vol 00:10:16

    That result feels closer to daily agent use than a single benchmark issue. People don't ask an agent to solve one frozen task and then throw the repo away. They ask for a feature, then a rename, then a constraint, then a test fix, then a migration, then another edge case. The agent needs stamina, but stamina here isn't vibes. It is the ability to keep the reference state aligned with the code after the fifth change.

26. Lenar Kess 00:10:43

    The predictive-validity paper is more of a methodology argument. It says aggregate leaderboards underspecify deployed agent evaluation, and it proposes ranking configurations by how well in-sample rank predicts out-of-sample rank. It cites a public-to-hidden competition result where execution-track rank correlation was negative 0.13, statistically indistinguishable from zero, while planning-track correlation was 0.69. The authors are open that the evidence is partial, but the proposed test is concrete.

27. Damra Vol 00:11:17

    That is the benchmark version of a lesson builders learn the hard way. A score is only as helpful as the shift it survives. If the rank order changes when the hidden cases arrive, then the leaderboard told you something about the public set, not about the agent you should trust.

28. Lenar Kess 00:11:33

    AutoPass makes the same point in a narrow compiler domain. It uses a multi-agent workflow for LLVM pass tuning. The agents inspect compiler internals, optimization remarks, intermediate representation, and runtime measurements, then refine pass pipelines under a small iteration budget. The reported geometric-mean speedups over LLVM O3 are 1.043 on x86-64 and 1.117 on ARM64. Small numbers, but compiler optimization lives on small numbers that repeat forever.

29. Damra Vol 00:12:11

    And it has the important rollback rule: the system keeps the best validated pipeline and falls back to O3 if the candidate doesn't beat it. That sounds obvious until you remember how many agent demos

reward novelty over regression control. In compiler tuning, the machine gets to answer back. It either runs faster on the target or it doesn't.

30. Lenar Kess 00:12:33

    So the agent story today is pretty grounded. Coordination before the pull request, stamina across turns, rank stability beyond the public benchmark, and runtime evidence for optimization. None of that is as flashy as a model release. It is closer to the work that decides whether model releases become dependable tools.

31. Damra Vol 00:12:53

    I'd add one caveat. The papers are still papers. grite's cleanest numbers are synthetic. StaminaBench uses generated REST APIs, which is a good interface but not the full mess of product code. Predictive validity is a proposal with a pilot design. AutoPass depends on a constrained compiler setting. But the shared pressure is easy to recognize: agents need evidence they can act on, not just bigger prompts.

32. Lenar Kess 00:13:20

    Startup Fortune reports that Hyundai is buying SoftBank's remaining 9.65 percent stake in Boston Dynamics for $325 million, which would give Hyundai full ownership of the robotics company. The article says Hyundai is expected to approve the purchase on Monday, June twenty-second, and places it against the earlier 2021 deal where Hyundai bought control of Boston Dynamics.

33. Damra Vol 00:13:46

    I'd keep one caution in view: this is one report, and the company filing or announcement would make the deal details firmer. But the ownership logic is clear enough to discuss. SoftBank exits a slower product-company bet. Hyundai consolidates a robotics asset that can be tested inside its own factories. That changes the patience model.

34. Lenar Kess 00:14:08

    Exactly. Boston Dynamics has always had this strange dual identity: astonishing demos, long commercial timelines. Spot became the practical product. Atlas is the hard one because humanoid robots have to compete with existing factory automation, not just with other humanoids. The Startup Fortune piece points to Atlas beginning work at Hyundai's Georgia electric vehicle plant by 2028, starting with parts sequencing and moving toward harder operations later.

35. Damra Vol 00:14:36

Factory ownership matters there because the first customer isn't imaginary. Hyundai can choose the task, control the environment, build the maintenance loop, and measure whether the robot improves production. A robotics company selling into a random factory has to survive procurement, integration, safety review, and the local reality of a floor it doesn't own. Hyundai can make Boston Dynamics part of the operating environment.

36. Lenar Kess 00:15:03

    The article also quotes the reliability bar that Boston Dynamics CEO Robert Playter has talked about elsewhere: Atlas would need to learn new factory tasks in a day or two and reach 99.9 percent reliability before it is truly useful on the floor. That number is almost comically unforgiving until you imagine a robot stopping a line, damaging a part, or needing a human nearby for every correction.

37. Damra Vol 00:15:31

    It is unforgiving because the factory is already optimized around machines that do specific jobs very well. A humanoid has to justify its generality. It has to be flexible enough to beat fixed automation on task variety, and reliable enough that the flexibility doesn't become babysitting.

38. Lenar Kess 00:15:50

    SoftBank's exit also makes sense in that light. The article describes SoftBank moving capital toward larger AI infrastructure bets. Boston Dynamics is physical, slow, and product-bound. Hyundai's bet is narrower but more measurable: make the robot valuable in Hyundai plants first. If that works, the robotics platform has a reference customer with real work behind it.

39. Damra Vol 00:16:14

    And then the robotics research papers today give the technical companion to the ownership story. Ownership decides who gets to run the loop. The papers ask what the loop has to contain. Resets, verification, grasp data, real-world feedback, and a way to make the robot learn without a human sitting there turning every trial into a bespoke experiment.

40. Lenar Kess 00:16:36

    The ENPIRE paper from NVIDIA, Carnegie Mellon, and Berkeley describes an agent harness for real-world robot policy self-improvement. The abstraction is refreshingly concrete: reset the scene, execute a policy, verify the outcome, and refine the next iteration. The authors split that into environment construction, policy improvement, rollout, and evolution across multiple robots.

41. Damra Vol 00:17:01

That reset step isn't a footnote. In robotics, the experiment is physical. If the robot knocks the object somewhere weird, bends the zip tie, shifts the pin box, or gets stuck halfway through a task, the next trial is polluted unless the environment is restored. ENPIRE's claim is that coding agents can help build the reset and verification APIs, then use them as stable interfaces for policy improvement.

42. Lenar Kess 00:17:29

    The paper reports frontier coding agents autonomously developing policies that reach 99 percent success on tasks like Push-T, organizing pins into a pin box, and cutting a zip tie. It also has a fleet angle: eight bimanual robot stations, one agent per station, with agents testing hypotheses asynchronously and sharing successful recipes. Scaling from one to eight agents reduced time to target success in their Push-T and pin-insertion experiments, but token use grew faster than the ideal linear trend at eight agents.

43. Damra Vol 00:18:02

    That is a very physical version of the budget story. More robots can shorten wall-clock time, but the agent team spends tokens reading logs, writing code, summarizing peer branches, and waiting on model calls. The scarce resource isn't only the robot. It is the coordination overhead between the robot, the GPU, the model, and the experiment history.

44. Lenar Kess 00:18:26

    The Human Universal Grasping paper takes a different route. HUG trains on human grasp data collected with smart glasses: one million egocentric frames, 6,707 object recordings, across 41 buildings. The model takes an RGB-D image and a clicked object point, predicts a human hand grasp, and retargets it to robot hands. It is trained on human data, not robot demonstrations.

45. Damra Vol 00:18:55

    That is a lovely idea because humans are the data source that already exists at absurd scale. We pick up cups, pens, brushes, boxes, and weird household objects all day. The trick is turning that into metric, retargetable, robot-usable grasp information rather than video vibes.

46. Lenar Kess 00:19:14

    Their benchmark is also concrete. HUG-Bench has 90 unseen objects across five geometric categories and different sizes. In real-world tabletop experiments on 30 test objects, HUG reaches 66.7 percent success, compared with 43.7 percent for Dex1B and 32.7 percent for Contact-Anchored Policies. The paper says HUG grasped 28 of the 30 objects at least once. Those aren't magic numbers, but they are useful numbers.

47. Damra Vol 00:19:48

And the failures are informative. Some objects are too large or awkward for the hand. RGB alone performs badly. Point cloud alone gets close but can grab the wrong part of an object. The combined RGB and point-cloud model works better because the robot needs both geometry and semantics. That is the whole robotics story in miniature: the body, the sensor, the object, and the policy all get a vote.

48. Lenar Kess 00:20:15

    Put those two papers beside Hyundai and Boston Dynamics, and the robotics question becomes less theatrical. Who owns the place where the robot learns? Who controls reset, verification, repair, parts, and uptime? The research says feedback loops are becoming more automatable. The ownership news says the first valuable loop may live inside a factory that already belongs to the robot's parent company.

49. Damra Vol 00:20:41

    That is also why I would keep the humanoid race at human scale. Tesla, Figure, Unitree, Boston Dynamics: the names are easy to list. The harder work is deciding which tasks let a general robot earn its keep before it becomes general. Hyundai's advantage isn't that Atlas looks better in a video. It is that Hyundai can give Atlas a constrained job and keep iterating until the reliability number is less embarrassing.

50. Lenar Kess 00:21:08

    There were also two Reddit reposts today about John Jumper leaving Google DeepMind for Anthropic, plus a Susan Zhang X post in the source list about China-risk discourse and talent restrictions. I'm not going to make that a main story here, because Construct covered the Jumper move deeply yesterday and today's sources don't add a new primary statement.

51. Damra Vol 00:21:30

    That altitude is right. Jumper moving to Anthropic is a major talent event, especially because AlphaFold wasn't just a model success but a whole scientific workflow. But a duplicate Reddit gallery doesn't become new evidence because it appears in two communities. And the Susan Zhang item needs the full X context before we build a strong claim on it.

52. Lenar Kess 00:21:52

    The broader policy surface is still alive, though. Over the last week, we have had model access restrictions, export-control anxiety, and now talent movement getting read through China-risk arguments. The practical question is who gets to hire, move, collaborate, and be trusted around frontier systems. But today's sources support only a marker, not a full segment.

53. Damra Vol 00:22:15

    And the marker is useful precisely because it tells us what evidence would change the story: a lab statement, an employment policy, a visa rule, a government filing, or a researcher explaining why they moved. Without that, the discourse can outrun the artifact very quickly.

54. Lenar Kess 00:22:32

    So that's today's map. DeepSeek is trying to make million-token context efficient enough to use. Companies are discovering that model usage needs a meter. Agent researchers are measuring coordination, stamina, and evidence. Robotics is moving from demos toward owned feedback loops. The next useful signal isn't a bigger claim. It is a system that can show its work, pay its bill, and improve from the feedback it actually receives. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic