DISPATCH 096 · 2026-07-24 | GSV PERMISSION WAS REMEMBERED

# Your Medical Record Enters the Conversation

2026-07-24 / 00:18:38

> *“A permission prompt can govern a single answer. It has a harder time governing what a person gradually comes to believe.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

ChatGPT can now carry medical records and Apple Health data into ordinary conversations, which makes personal context more useful and the consequences of a mistaken answer more intimate.

- OpenAI's Health announcement explains the U.S. rollout, per-response permissions, connected records, physician evaluations, and the company's claim that more than 300 million people ask ChatGPT health questions each week.
- Mike Takahashi's AgentForger disclosure shows how one crafted link could create and schedule a Workspace Agent with access to previously authorized connectors; OpenAI fixed the flaw four days after disclosure.
- The Soofi S paper describes a German-and-English mixture-of-experts model trained on German infrastructure, with three billion of thirty billion parameters active per token and unusually extensive release commitments.
- Nathan Lambert's open-model dashboard note measures downloads and derivatives by model origin and reports that U.S. families remain well behind China and Qwen.

- Reuters on Alphabet's spending puts the AI buildout on the cash-flow statement: a $5.9 billion second-quarter burn arrived beside record cloud growth.
- Runway's Media Router announcement turns cost, quality, and latency preferences into model-selection policy, while Replit's hosting update promises a price cut of more than half for high-volume apps beginning August 1.

---

SEGMENTS



# Transcript

1. Lenar Kess 00:00:04

   OpenAI began rolling Health in ChatGPT out to U.S. adults on Thursday, and the new part is easy to picture. You connect Apple Health or a supported medical record, ask about a lab result, and ChatGPT can compare it with earlier tests instead of making you paste six months of history into a box. The company says more than 300 million people already ask ChatGPT health questions each week. An existing habit now gains access to medications and visits. It can also draw on sleep, activity, and the awkwardly complete archive that lives across patient portals.

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

- x.com
- x.com
- x.com
- reuters.com
- x.com
- x.com
- x.com
- x.com
- openai.com
- openai.com
- labs.zenity.io
- investing.com
- substack.com

2. Damra Vol 00:00:41

   OpenAI says seventy percent of health conversations among early users happened outside the dedicated Health area. That detail changes the product for me. People don't encounter health as a separate mode. A restaurant search touches an allergy, a weekend plan touches a recent injury, and a recipe touches a medication or a dietary restriction. Health data moves into the rest of the assistant because the person already lives that way.

3. Lenar Kess 00:01:08

   Right. The announcement says the assistant can use connected health information across ordinary conversations, with permission. By default it asks before using a medical record or Apple Health data for a response. You can approve that request once, or choose always allow and stop seeing the prompts. Connected records and conversations that use them aren't used to train OpenAI's foundation models or target ads, according to the company.

4. Damra Vol 00:01:33

   That permission design is sensible, and it still leaves a human problem. The first request may be obvious: yes, use my lab results to explain this number. The fiftieth request may be restaurant planning where an old diagnosis changes the recommendation in a way I didn't anticipate. A permission prompt can govern a single answer. It has a harder time governing what a person gradually comes to believe because the assistant has been carrying a medical biography through otherwise ordinary chats.

5. Lenar Kess 00:02:04

OpenAI draws one boundary there. Memories may be created from Health conversations, but the announcement says memories aren't created directly from the connected records or Apple Health feed. If you disconnect a source, synced data is deleted from OpenAI's systems within thirty days; information already written into conversation history stays until you delete those conversations. The privacy notice also says a limited number of authorized staff and service providers may access Health data for safety improvement unless the user opts out.

6. Damra Vol 00:02:37

   That leaves three different objects in a person's head: the source record, the conversation that used it, and the memory formed from what the person said about it. Those objects have different deletion and access rules. Most people won't maintain that mental model. They'll remember that they connected Health, and they'll treat the rest as one container.

7. Lenar Kess 00:02:57

   OpenAI's examples include comparing a new result with prior tests, summarizing changes since an appointment, explaining a visit note in plain language, and preparing questions for a follow-up. One early tester describes turning diagnoses, imaging results, and surgery notes into a timeline they could explain to a physical therapist. I understand the appeal. Medical records are often written for billing, handoff, and liability before they're written for the patient who owns the body.

8. Damra Vol 00:03:27

   And the assistant has time. A clinician may have twelve minutes and a record full of copied-forward text. The model can keep translating until the person understands the difference between a finding, a possibility, and a decision. That doesn't establish clinical efficacy. It does address the moment after a visit when you remember half a sentence, open a portal, and find four pages of abbreviations.

9. Lenar Kess 00:03:52

   OpenAI says hundreds of physicians helped create scenarios and scoring rubrics. They assessed accuracy and safety. The review also covered communication, context awareness, completeness, and appropriate escalation. The company says every GPT-5.6 model beat GPT-5.5 on HealthBench Professional, and physicians tested the connected-data experience before release. Those are company-reported evaluations. We don't have an independent assessment of this rollout in today's sources, and access to more context doesn't prove that the answer will be medically sound.

10. Damra Vol 00:04:28

    [pause] More context can make a wrong answer more persuasive. A generic mistake has friction because it sounds generic. A mistake that mentions your prescription history, last winter's lab result, and Tuesday's sleep data arrives wearing your life as evidence. The model may be better

because it has the record. The person may also grant it more authority than the measured improvement deserves.

11. Lenar Kess 00:04:51

    OpenAI explicitly says the product supports rather than replaces professional care, and its examples are mostly explanation and preparation. Yet the interaction is conversational, personalized, and available at the exact hour when a person is anxious. Product language can draw a boundary around diagnosis. Human behavior will test that boundary every night.

12. Damra Vol 00:05:13

    The adoption number makes that test enormous. Three hundred million weekly health questioners aren't waiting for a medical chatbot category to mature. They're already asking. OpenAI is reducing the effort required to give those conversations a history. If this works well, people arrive at appointments better prepared. If it works badly, the error gets woven into weeks of follow-up questions before a clinician ever sees it.

13. Lenar Kess 00:05:38

    Early rollout data should answer three narrower questions: how often does the model ask for missing context, how often does it recommend urgent care appropriately, and how often do users correct stale records? Those rates would tell us whether connected context helps the conversation stay anchored, or merely gives the answer more detail.

14. Lenar Kess 00:05:59

    Security researcher Mike Takahashi published AgentForger on Thursday, a flaw in ChatGPT's Workspace Agent builder that OpenAI had already fixed in June. The claimed attack began with one link. A logged-in employee with access to Workspace Agents and at least one previously authorized connector could click it, and the builder would start creating an agent from instructions embedded in the address.

15. Damra Vol 00:06:25

    The odd detail is that the address parameter wasn't filling the prompt box for review. Takahashi found that it was submitted and executed when the page loaded. The attacker could select the chief-of-staff template and provide the first instruction to the builder in the same URL. The victim supplied the authenticated session simply by being logged in.

16. Lenar Kess 00:06:46

His proof of concept told the builder to attach existing connectors, switch approval settings for write actions from always ask to never ask, publish the agent, and install recurring schedules. It then instructed the agent to read emails from the attacker, execute tasks from those messages, and email results back. No new authorization screen appeared because the victim had connected services such as Outlook, Gmail, Slack, Google Drive, SharePoint, or Teams before the click.

17. Damra Vol 00:07:15

    The phish didn't steal a password. It borrowed completed consent. That is nastier in a very specific way: the connector authorization was legitimate, the user session was legitimate, and the builder was an official OpenAI surface. The malicious part was the instruction that assembled those legitimate pieces into a new operator.

18. Lenar Kess 00:07:36

    Takahashi says the builder worked through the request, disabled the approval gate, published the agent, and used Preview Mode to execute it immediately. The schedules then woke it every five minutes to check for new instructions. He describes the combination as untrusted input from the URL, private data through connectors, and an outbound path through email.

19. Damra Vol 00:07:57

    Preview Mode is a lovely institutional name for a process that touched live accounts. [tsk] A preview that sends mail or reads connected data is an execution mode with a reassuring label. The exploit made that mismatch visible because it crossed from configuration into action before the employee got another chance to decide.

20. Lenar Kess 00:08:17

    The disclosure timeline matters. Zenity reported the issue through Bugcrowd on June fourth. Bugcrowd triaged it and OpenAI accepted it on June fifth. OpenAI fixed it on June eighth, four days after the original report. There is no claim here of exploitation in the wild, and the published demonstration had prerequisites. This was a disclosed and repaired vulnerability, not a notice that every Workspace Agent had been taken over.

21. Damra Vol 00:08:45

    The repair doesn't erase the design lesson. Natural language was allowed to change approval policy and scheduling while the same flow could execute against authorized services. An agent builder is partly a programming environment and partly an identity console. If a sentence can lower the approval level, create persistence, and launch a run, then sentence interpretation sits inside the security model.

22. Lenar Kess 00:09:10

    On Tuesday, we talked broadly about production-agent security. This disclosure earns separate treatment because it gives us an exact path: initialization state arrived through a link, the prompt changed security-sensitive configuration, Preview touched connected accounts, and the scheduler created persistence. Four ordinary product features composed into the exploit. OpenAI's fast fix closed this route; other agent builders still have to decide where conversational convenience ends and an explicit authenticated action begins.

23. Lenar Kess 00:09:44

    The Soofi team revised its paper this week for Soofi S, a German-and-English open foundation model built on Deutsche Telekom's German Industrial AI Cloud in Munich. It is a mixture-of-experts hybrid Mamba Transformer with thirty billion parameters total and three billion active for each token. The team says the architecture keeps its key-value cache nearly constant as the context grows, which targets long conversations and many simultaneous users.

24. Damra Vol 00:10:14

    That is a sovereign-model claim with an artifact attached. The model was trained on roughly twenty-seven trillion tokens, with German deliberately weighted more heavily. The paper promises the weights and selected intermediate checkpoints. It also promises per-source data accounting, hyperparameters, and the code used for training and evaluation. Commercially licensed sources get aggregate statistics and exact mixture accounting where the raw material can't be redistributed.

25. Lenar Kess 00:10:43

    The benchmark claims are the authors' own. They say Soofi matches dense models in the fourteen-to-twenty-seven-billion-parameter range across aggregate German and English evaluations, leads the code aggregate in both languages among seventeen open base models, and beats the European sovereign baselines in their comparison. This is version three of a paper first submitted July tenth, not the model's first appearance today.

26. Damra Vol 00:11:09

    You can test a sovereignty claim by listing the nouns. Where did the training run? Who can inspect the data mixture? Can another lab reproduce the recipe? Which legal permissions travel with the weights? Soofi answers more of those than a flag attached to a download page. Capability still needs outside testing, especially in German, but the release commitment is substantial.

27. Lenar Kess 00:11:32

Nathan Lambert's team also published a dashboard that updates daily and compares open-model downloads and derivative models by country and organization. Lambert's summary says the U.S. role is growing slowly and remains well behind China and Qwen. The dashboard is measuring adoption activity on Hugging Face, not intelligence, revenue, or the nationality of every person using a model.

28. Damra Vol 00:11:56

    Downloads and derivatives answer different questions too. A download can come from a benchmark farm or one company pulling the same weights across a fleet. A derivative says somebody invested enough to fine-tune or adapt a family, but it doesn't tell you whether the result has users. The dashboard makes the ecosystem visible, and its categories still need to be read as proxies.

29. Lenar Kess 00:12:19

    Lambert's policy argument is that American open-model development needs much more support, and he points to Qwen's lead as evidence. The measurement tells us more than the slogan. On Tuesday, the story was a speculative American ban on Chinese open weights that disappeared within hours. Today's artifacts let us ask a better question: which model families are people choosing to extend when nobody forces the choice?

30. Damra Vol 00:12:46

    Soofi also complicates the country race. Its training infrastructure is German, its languages are German and English, its architecture draws from work developed across the field, and its users may be anywhere. A country can own the compute and release process while the downstream ecosystem forms somewhere else. Sovereignty describes control over a set of dependencies; adoption describes what other people decide to build.

31. Lenar Kess 00:13:13

    Outside groups now need to run the Soofi weights, check the German evaluations, and measure the claimed long-context throughput. The dashboard can then show whether a well-documented European model becomes a family that others extend, or remains an impressive national release with little downstream activity.

32. Lenar Kess 00:13:31

    Reuters reports that Alphabet burned five-point-nine billion dollars in the second quarter, its first cash burn on record. Google Cloud grew eighty-two percent during the same period. The company now expects to spend fifteen billion dollars more in 2026 than it previously planned, and it predicts another increase next year.

33. Damra Vol 00:13:52

    Those two facts belong beside each other. Cloud demand is surging, and serving that demand consumes more cash than Alphabet is generating in the period. Reuters says Alphabet plans to rent additional data-center capacity from other companies even though doing so will reduce margins. Spending is rising while the product has plenty of customers. Capacity is constrained, and the financial return trails the construction bill.

34. Lenar Kess 00:14:18

    Reuters estimates that the largest technology companies will spend more than seven hundred billion dollars this year, with debt and share sales covering more of the gap. For Alphabet, capital expenditure is expected to reach forty-one percent of revenue, up from twenty-three percent. That ratio gives the investor anxiety a concrete basis: a much larger portion of each sales dollar goes back into data centers and equipment.

35. Damra Vol 00:14:45

    One negative cash-flow quarter can't settle the value of the whole buildout. At least twenty brokerages raised their Alphabet price targets after the same results, according to Reuters, because the cloud business grew so fast. Investors can believe demand is excellent and still dislike the amount of cash required to meet it. Those positions fit together perfectly well.

36. Lenar Kess 00:15:07

    The competitive pressure makes restraint expensive. If Microsoft, Amazon, Meta, or Alphabet slows construction while customers are waiting for capacity, the customer may move. If all of them keep building, more compute eventually becomes interchangeable and returns can fall. Reuters quotes market strategist Lale Akoner making that second point: providers may spend more while accepting lower returns as models get cheaper and capacity becomes easier to obtain.

37. Damra Vol 00:15:37

    Next week's reports from Microsoft, Meta, and Amazon will show whether Alphabet is an outlier or simply first through the earnings door. The comparable numbers are spending as a share of revenue and cash generated after investment, alongside cloud growth and any new capacity leases. A stock-price reaction alone can't tell you which part investors objected to.

38. Lenar Kess 00:15:59

    Runway introduced Media Router on Thursday. A developer states a preference for cost, quality, or latency, and the router chooses among generative-media models rather than making the developer

select one model for every request. Runway and its chief executive, Cristóbal Valenzuela, present it as a new control layer for media generation.

39. Damra Vol 00:16:21

    The router needs to reveal what happened after it makes a choice. If today's request goes to a fast model and tomorrow's goes to a better one, can you inspect why, reproduce the decision, and compare the outputs? The announcement doesn't include a benchmark or price comparison, so we can't say the policy produces better media.

40. Lenar Kess 00:16:40

    I like the abstraction because creative work already has variable intent. A storyboard thumbnail, a client preview, and a final plate don't need the same model. A preference can be more natural than a product name. The router becomes much more valuable if it records enough evidence for a creator to understand why the visual character or bill changed between runs.

41. Damra Vol 00:17:02

    And it makes Runway's taste part of the service. Cost and latency can be measured directly. Quality depends on the prompt, the image, and the person deciding whether the result works. A preference-optimized router has an evaluation system somewhere inside it, even if the announcement doesn't show us that system yet.

42. Lenar Kess 00:17:21

    Replit says hosting prices for high-volume apps will fall by more than half on August first. Amjad Masad says the company can now negotiate better cloud-provider discounts because it buys more infrastructure and is passing those savings through to customers. We don't have a before-and-after price table in today's sources, so the claim stays at that level.

43. Damra Vol 00:17:42

    The explanation is ordinary purchasing power. Replit aggregates many small application workloads, buys infrastructure on better terms, and lowers the meter for the people above it. The reduction applies to high-volume apps, rather than every kind of Replit usage, and existing users will be able to compare their own bill after August first.

44. Lenar Kess 00:18:04

    Replit's update gives us a concrete contrast with Alphabet. At one end of the market, a hyperscaler is spending so fast that cloud growth and cash burn appear in the same quarter. At the other, a software platform says its larger purchase volume has become a customer price cut. Those are

different businesses and different accounting stories. Both claims will soon produce numbers we can check: Alphabet will report another quarter, and Replit users will receive an August bill. Lenar Kess.

## Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic