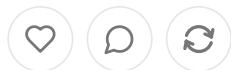


HHS Goes All-In on ChatGPT for State Audits: What the May 2026 Generative AI Fraud Expansion and February's CRUSH RFI Mean for Payment Integrity, Provider Audit Defense, and the Health Tech Buy-Side

MAY 24, 2026



Share

Video Preview



 **Part I Podcast free on Spotify.**



Thoughts on Healthcare Markets & Technology

 **Part I: HHS Goes All-In on ChatGPT for State Audits: What the May 2026 Generative AI Fraud Expansion & Feb's CRUSH RFI Mean for Payment Integrity, Provider Audit Defense & the Health Tech Buy-Side**

HHS shipped an LLM workflow for federal audits without an RFP. No vendor circus. No multi-year authorization. That is the actual story from May 21...

▶ Listen now

an hour ago · Thoughts on Healthcare

Part II Podcast episode for subscribers only. Also on Spotify.



Thoughts on Healthcare Markets & Technology



Part II: HHS Goes All-In on Audits: What the May 2022 Expansion & Feb's CRUSH Integrity, Provider Audit D Buy-Side

HHS shipped an LLM workflow for federal audits without an RFP. No vendor circus. No multi-year authorization. That is the actual story from May 21...

▶ Listen now

an hour ago · Thoughts on Healthcare



Discover more from Thoughts on Healthcare Markets & Technology

Expert analysis of healthcare and life sciences markets, technology, investment, entrepreneurship policy, and AI — for investors, entrepreneurs,...

Over 3,000 subscribers

Enter your email...

Subscribe

By subscribing, you agree Substack's [Terms of Use](#), and acknowledge its [Information Collection Notice](#) and [Privacy Policy](#).

Already have an account? [Sign in](#)

To listen to paid episodes in Apple or Spotify, link your Substack subscription via the show settings on those platforms (instructions inside the Substack app under Subscriptions → Podcast).

Thoughts on Healthcare Markets & Technology is a reader-supported publication. To receive new posts and support my work, consider becoming a free or paid subscriber.

Table of Contents

1. What HHS actually did on May 21, and what got lost in the press cycle
2. The CRUSH RFI was the real signal back in February
3. Who built this, and who is going to build the next version
4. The RFP question, the procurement vehicles, and why there is no single mega contract
5. What the government can do now versus what industry still has to build
6. Payment integrity vendors, the incumbents, and the new provider audit defense category
7. The investor angle, with three buckets of capital deployment
8. Selling to the feds as a health tech founder, a casual playbook
9. Closing skepticism, the legal exposure, and the next 18 months

Abstract

- On May 21, 2026, HHS Asst Sec for Financial Resources Gustav Chiarello announced a rolling generative AI ingestion of state and grantee single audits from any entity spending >\$1M/yr in federal funds. Tools include ChatGPT plus other LLMs. Targets: chronic noncompliance, repeat deficiencies, material weaknesses, delinquent audit obligations. Penalty: loss of funding.
- Builds directly on the Feb 27, 2026 CMS CRUSH RFI (Comprehensive Regulations to Uncover Suspicious Healthcare), comments closed 3/30/26. Topics: enrollment screening, identity proofing, ownership controls, MA preclusion, lab fraud (genetic + molecular dx), DMEPOS reform, claim filing deadline compression, AI-assisted coding review, beneficiary solicitation, surety bonds, Medicaid integrity, Marketplace integrity (FFE + SBE).
- 2025 enforcement baseline CMS cited: \$5.7B Medicare payments suspended, 122,658 claims denied, 5,586 billing privileges revoked, 372 fraud referrals worth \$3.7B. Concurrent actions: nat'l 6-mo DMEPOS moratorium, \$259.5M Medicaid funds withheld from MN.
- Healthcare payment integrity market: \$15.12B (2025) to \$28.02B by 2030 at ~13.1% CAGR per Mordor. 6.26% claim error rate persisting. 60%+ claims now cloud-processed. Roughly \$20B in legacy mainframe admin overhead still being retired.
- Gov-side incumbents: Cotiviti GOV (RAC Regions 3, 4, 5), Optum FWA, Performant, Qlarant, Gainwell (post HMS), SAS Institute, NTT Data. Commercial-side incumbents: Cotiviti, Optum, Codoxo, EXL/SCIO, ClarisHealth, Healthcare Fraud Shield, Shift Tech, Lyric, Zelis, HealthEdge, Sagility, Multiplan.

- Oz quote: “padlocking the jar and letting them starve.” RFK Jr. framing: “pay and chase” to “detect and deploy.”
- Three layers of buy-side opportunity emerging: (1) gov-side document-heavy unstructured data tooling, (2) provider-side audit defense and counter-AI, (3) infra (eval, observability, explainability, fed-compliant RAG) underneath both sides.

What HHS actually did on May 21, and what got lost in the press cycle

The press cycle made it sound like HHS pointed Skynet at every Medicaid claim. The actual scope is narrower, more interesting, and more strategically important than the headlines.

Thoughts on Healthcare Markets &
Technology is a reader-supported
publication. To receive new posts and
support my work, consider becoming a
free or paid subscriber.

Single audits are filed under OMB Uniform Guidance, 2 CFR 200 Subpart F, by any entity that spends \$1M or more of federal money in a year. States, counties, public universities, hospital systems with grant funding, nonprofits running addiction services or research. Audits land at the Federal Audit Clearinghouse. Most then sit there. Chiarello, who is the Assistant Secretary for Financial Resources (not exactly a household name), put it bluntly in the AP interview: everyone files an audit, it lands with a thud, nobody does anything. HHS is now running those PDFs through LLMs to surface chronic deficiencies, then sending formal letters that say, in effect, fix this or lose your funding. Chiarello also told reporters other federal departments could pretty easily copy the approach.

That framing matters for at least three reasons. First, the analysis is happening on already-public reports, which dodges most of the PHI and data-sharing issues that have bogged down federal AI projects for a decade. Second, HHS did not need new statutory authority to do this. The agency was sitting on the documents already. Third, it sets the political and procurement template for the next, much harder thing, which is bringing AI into pre-pay claims integrity on actual Medicaid and MA data flows.

The press reaction underplayed all three of those points and overplayed the dystopian framing. Anyone tracking the space should ignore the surface narrative and pay

attention to the fact that HHS just shipped its first internal LLM workflow without an RFP, without a vendor selection circus, and without lawyers having a stroke. That is unprecedented for this agency at this scale.

The CRUSH RFI was the real signal back in February

Three months before the May announcement, CMS dropped the CRUSH RFI in the Federal Register on Feb 27, 2026, with a 30-day comment window that closed March 30. CRUSH stands for Comprehensive Regulations to Uncover Suspicious Healthcare, which is exactly the kind of acronym a brand consultant would have flagged but apparently nobody at CMS asked.

The RFI is enormous. It covers more than a dozen subject areas, each of which is its own enterprise software market. The biggest themes: enhanced provider enrollment screening with identity proofing and stronger ownership controls including potential citizenship and residency tests on beneficial owners. Preclusion list reform to stop revoked providers from billing Medicare Advantage. A possible requirement that MA providers also enroll in traditional Medicare so CMS has a continuous view. New tools against laboratory test fraud, particularly genetic and molecular diagnostics where the FWA economics have gone berserk over the last five years. DMEPOS supplier overhaul including the now-active six-month national moratorium on new enrollment for prosthetics, orthotics, registered pharmacist, and respiratory therapist supplier categories. Tighter claim filing deadlines, possibly compressing the current one-year window. AI-assisted coding accuracy and medical record review. Expansion of the DMEPOS telephone solicitation ban to other channels. Higher surety bonds and broader provider types subject to them. Stronger Medicaid and CHIP integrity tools, with more state authority and incentive payments. Marketplace integrity for both the federally facilitated exchange and state-based exchanges, with one of the literal questions in the RFI asking how CMS could use advanced technologies including AI to prevent, detect, and address FWA in those exchanges.

If a stakeholder read that menu and did not file comments, they will be eating whatever CMS decides on their behalf in the proposed rulemaking that follows. The Hall Render, Foley Hoag, DLA Piper, and Applied Policy memos all flagged how aggressive the program could become. The AHA filed its comments on March 30 and pushed back on automated downcoding, hallucination risk in AI coding tools, and the absence of independent physician review of coverage denials. AHA also pre-positioned legal arguments that will show up in court two years from now when the first FCA actions hit.

Mehmet Oz, in his administrator capacity, gave the line of the year for this announcement: CMS is done trying to catch fraudsters with their hands in the cookie jar, instead they are padlocking the jar and letting them starve. Whatever one thinks of the metaphor, the policy is real. Kennedy, as HHS Secretary, framed the shift as moving from pay and chase to detect and deploy. That phrasing is not a casual rebrand. It is a specific procurement and architecture signal. Pay and chase is recovery audit contractors, retrospective review, and recoupment workflows. Detect and deploy is pre-pay analytics, real-time scoring, and automated suspension at the EFT layer. Different vendors, different tech stack, different unit economics.

Who built this, and who is going to build the next version

For the May 21 announcement specifically, HHS appears to be using off-the-shelf ChatGPT and other unnamed tools. There was no public RFP for the single audit ingestion work. Chiarello told reporters the program is being run inside his office with existing tooling. This is not a flex on the technology, it is a flex on procurement. HHS effectively bypassed the entire federal acquisition stack by using commercial AI products on already-public data.

The previous generation of HHS AI work, going back to 2023, used tree-based machine learning models for outlier billing in Medicare. The then-HHS CIO Karl Mathias was on the record at FedScoop confirming that work. Those tree models are not glamorous, but they consistently beat newer architectures on tabular claims data, which is why the payment integrity incumbents still ride them. Beneath the gen AI announcement is a much older statistical fraud detection infrastructure that has been in production for years and will continue to be the workhorse.

For the next version, several procurement paths matter. Cotiviti GOV Services already holds the CMS Recovery Audit Contractor contracts for Regions 3, 4, and 5, awarded last year through competitive procurement. That gives Cotiviti structural advantage on any task order modification or scope expansion to pre-pay or AI-assisted review. Performant and Qlarant still hold UPIC, MEDIC, and SMRC work for various regions. Gainwell controls a big chunk of state Medicaid MMIS infrastructure since acquiring HMS, which historically did Medicaid third-party liability and credit balance recovery. SAS Institute is embedded in the Center for Program Integrity analytics shop. Palantir Foundry already sits inside CMS through the broader DOGE-era data integration push. Microsoft Azure Government, particularly the Secret-cleared Azure OpenAI Service from the August 2024 partnership with Palantir, is the most likely hosting environment for anything classified or PHI-handling.

So when somebody asks who built this, the honest answer is: ChatGPT for the audit summarization, SAS and tree-based incumbents for the Medicare claims work, Palantir Foundry for data plumbing, and a long tail of payment integrity contractors for execution. None of them got a fresh contract for the May announcement. All of them are positioning for the next one.

The RFP question, the procurement vehicles, and why there is no single mega contract

A common mistake when reading announcements like this is assuming there is a giant capstone RFP coming. There almost certainly is not. CMS and HHS run procurement through a sprawl of vehicles, and the AI work will get spread across them rather than concentrated in a single award.

The biggest vehicles to watch: CMS' SPARC (Strategic Partners Acquisition Readiness Contract), the agency's main IT and analytics ID/IQ. NIH and CDC use NITAAC CIO-SP4 for IT and a portion of analytics. GSA MAS (the consolidated multiple award schedule) handles most commercial software. The VA runs T4NG-2 for IT and is increasingly buying generative AI tools through OT authority. DHA has its own enterprise vehicles. For smaller, faster work, the OT authority (Other Transactions) lets agencies skip most FAR rules entirely, which is how DoD has been doing its AI buys. Phase III SBIR conversions can also be used to sole-source if the work came out of a prior SBIR award.

The CRUSH-derived rulemaking, when it lands, will probably spawn modifications to existing payment integrity contracts (Cotiviti, Performant, Qlarant, Gainwell) plus new task orders for things like beneficiary identity verification, AI-assisted coding review, and surety bond compliance. Each will be modest in dollar terms (tens of millions, not hundreds), but in aggregate the program will move billions over five years.

For a founder thinking about going after the work, the actual answer is not "respond to the RFP." It is: get on a vehicle, sub to a prime, build past performance, then prime your own task order on the second pass. The vehicle math typically takes 12 to 24 months. The FedRAMP authorization timeline is 12 to 24 months on top of that. Anyone who is not already on a vehicle and does not have FedRAMP Moderate is two to three years away from being able to invoice CMS at scale, regardless of how good their software is.

What the government can do now versus what industry still has to build

What the government has working in production today is a surprisingly limited capability set if you actually look at it. There is document summarization on text-heavy audits, courtesy of the May 21 announcement. There is tree-based ML on Medicare claims for outlier detection. There is statistical sampling for PERM and CERT improper payment measurement. There is Palantir Foundry doing cross-system entity resolution and data integration. There is human-in-the-loop investigation workflow at the UPICs and MEDICs.

What is missing, and where industry has to build, is a longer and more interesting list. Real-time pre-pay risk scoring at the EFT layer remains aspirational at the federal level despite Oz's padlock metaphor. Multimodal fraud detection that combines claims with medical records, telehealth video, audio of patient encounters, and prescription history barely exists in commercial product form. Provider identity continuity, meaning the ability to follow an individual provider through LLC layering, address farms, NPI changes, and ownership swaps, is poorly handled by every existing vendor including the incumbents. Explainable AI infrastructure that produces an evidentiary record sufficient for False Claims Act litigation is a serious gap. Eval and observability tooling for production AI in regulated workflows is still being built. The NY Medicaid data error that the administration had to publicly acknowledge to the AP is exactly the kind of issue better eval infrastructure would have caught.

Counter-AI tooling on the provider side does not really exist as a category yet. If CMS deploys AI-assisted coding review at scale, providers will need software that simulates the government model's risk scoring, identifies likely flagged claims before submission, and produces appeal-ready documentation. Today that work is handled by humans inside RCM teams and consultants. It will be productized within 24 months.

Privacy-preserving computation for cross-state Medicaid data sharing is another wide-open gap. Confidential computing, federated learning, and synthetic data all get tossed around in CMS conference panels, but none of them is in production at scale on Medicaid data flows. The technical and political problems are both nontrivial.

Payment integrity vendors, the incumbents, and the new provider audit defense category

The healthcare payment integrity market hit roughly \$15.1B in 2025 and is forecast to push to \$28B by 2030 at a 13.1% CAGR per Mordor. Claim error rates have stuck stubbornly around 6.26%, which is the part of the iceberg every CRUSH-style announcement is trying to crack. Cloud-based deployment now processes more than 60% of claims, retiring something on the order of \$20B in legacy mainframe admin expense. The competitive landscape is concentrated at the top and fragmented at the bottom.

On the gov side, Cotiviti GOV is the heavyweight, with the new RAC Region 3, 4, and 5 awards reinforcing its position. Performant and Qlarant cover UPIC and MEDIC zones. Gainwell, since absorbing HMS, controls major state Medicaid TPL and credit balance work plus MMIS implementations in many states. NTT Data, EXL/SCIO, and SAS Institute hold longstanding analytics contracts. Booz Allen, Leidos, and GDIT play the integration and systems-engineering role across all of this.

On the commercial side, Cotiviti is also the leader, with Optum FWA right behind. Codoxo has carved out a real niche with Explainable AI branding in Medicaid and commercial plans. ClarisHealth, Healthcare Fraud Shield, Shift Technology, Lyric, Zelis, HealthEdge, Multiplan, and Sagility round out the next tier. Most of these vendors are some combination of private equity owned (Cotiviti at Veritas, EXL public), inside UnitedHealth, or otherwise spoken for. The greenfield is not in commercial payer payment integrity. That boat sailed when private equity rolled up the category between 2018 and 2023.

The actual greenfield is provider-side audit defense. As of mid-2026, no clear category leader exists. Several players are circling: Iodine Software on CDI and utilization, CodaMetrix on autonomous coding, Janus Health on revenue cycle workflow, Reveleer on risk adjustment, RhythmX AI on payer-provider claims dispute. None of them is yet built around the specific use case of “CMS or a commercial payer has flagged your claim, what do you do next.”

That category will get built. The Pieces Technologies Texas AG settlement on hallucination rate disclosure already created the regulatory groundwork. The Lokken v. UHC discovery battle in Minnesota established that plaintiffs can subpoena AI model internals when contesting denials. Provider-side audit defense will be sold to hospitals, large physician groups, MA risk-bearing entities, and eventually to ACOs and direct contracting entities. The pricing model will probably mirror what Cotiviti and Codoxo do but inverted, so contingent fees on the dollar value of denials successfully appealed.

The investor angle, with three buckets of capital deployment

For anyone allocating capital to this thesis, three roughly orthogonal bets line up.

Bucket one is incumbents on both sides of the gov-commercial line. Cotiviti, Optum FWA, Gainwell, SAS, Codoxo. None of these is cheap, none is going to be a venture return, and most are PE-owned anyway so the trade is either secondary or waiting for an eventual IPO window. The right reason to hold exposure here is that the underlying market grew 13% CAGR and the public-policy tailwind just got materially stronger.

Bucket two is provider-side audit defense, which is the actually-venture-investible piece. Founders here need to be deeply technical, fluent in CPT and ICD coding, comfortable with both 837/835 transaction sets and FHIR US Core profiles, and willing to live inside the workflow pain of a hospital CFO's audit response team. The category is unbranded, the buyers know they need something, and the incumbents on the payer side are about to weaponize AI in ways that create predictable enterprise demand. Expect three to five Series A rounds in this space over the next six months, possibly led by Bessemer, General Catalyst's healthcare team, a16z bio, and the usual late-stage names sniffing earlier. The dark-horse buyers will be the EHR vendors, who could either build this in-house or acquire to bolt onto Epic Payer Platform and Oracle Health.

Bucket three is infrastructure underneath both sides. Eval and observability tooling for healthcare AI workflows. Hallucination detection on clinical and claims documents. Privacy-preserving computation for cross-state Medicaid data sharing. Synthetic data for model training where actual PHI cannot move. Identity resolution and provider continuity infrastructure. These are unsexy plumbing bets that almost nobody is funding in healthcare specifically, even though the horizontal AI infra market has been pouring money into the same primitives. Whoever puts a healthcare wrapper on Patronus, Arize, Trulens, or Galileo and gets HIPAA plus SOC2 plus FedRAMP done first is going to own a real piece of the next decade. None of the current generalist AI infra companies has prioritized health, which is the gap.

For asset allocators thinking about thesis-level exposure, the rough decomposition would be 40% incumbents (for the macro tailwind), 35% provider-side audit defense (for the venture upside), and 25% infra (for the optionality). Adjust per personal risk tolerance and conviction on regulatory tempo.

Selling to the feds as a health tech founder, a casual playbook

A common founder failure mode is to think federal procurement is “just enterprise sales with more paperwork.” That is wrong in roughly the way thinking running a marathon is just walking with more steps is wrong.

The actual playbook, in compressed form. Pick the vehicle before pitching anyone. The federal customer cannot buy from a company that is not on a contract vehicle they can use. For health AI work, the main ones are CMS SPARC, NITAAC CIO-SP4, GSA MAS, T4NG-2 at the VA, and SEWP V at NASA (which everybody uses because it is fast). Get on at least one. If that takes a year, fine, that is the cost of entry.

Sub to a prime first. Booz Allen, Leidos, GDIT, ICF, and Maximus all carry the past performance and the security clearances needed to win the actual award. Sub through one of them, deliver on a few task orders, build the past performance, then prime your own work on the next recompetes. This is the canonical path and there is no shortcut.

FedRAMP authorization is non-optional for anything touching CMS or HHS data. The classic path is FedRAMP Moderate via an agency sponsor, which costs roughly \$750K to \$1.5M and takes 12 to 18 months. The new FedRAMP 20x reform is supposed to compress that, but as of late 2025 the new process is still being shaken out. NIST 800-53 controls, FISMA Moderate or High classification, and ATO (Authority to Operate) are all gates. The cheap workaround is StateRAMP for state Medicaid work, which is easier and lets a vendor build muscle before going federal.

Bring data, not slides. Federal customers respond to performance benchmarks, hallucination rates, false positive rates, sensitivity and specificity on representative datasets, and side-by-side comparisons against incumbents. A federal contracting officer who has been doing this for 15 years can smell a marketing deck from across the room. The Pieces Technologies Texas AG settlement, where the company allegedly overstated its hallucination rate, will be cited in every federal AI vendor due diligence package for the next three years. Founders should pre-empt that line of questioning by publishing their eval methodology.

Speak the language. Use phrases like task order, vehicle, ID/IQ, ATO, past performance, prime, sub, set-aside, GWAC, BPA, FAR Part 12, OTA, FedRAMP boundary, ConMon, and ATO inheritance. If those words do not roll off the tongue, hire someone for whom they do.

Pick the right entry point. CMS is the obvious target but also the slowest and most political. DHA (Defense Health Agency) and the VA are often faster and let a vendor

build product muscle on real clinical data. NIH and CDC have smaller AI budgets but easier procurements. HRSA and SAMHSA care a lot about behavioral health and addiction services, which intersects directly with the single audit work HHS just announced. The Indian Health Service has the loosest procurement and the most operational pain.

Founders also need to understand the political tempo. The current administration's stated priorities are fraud, identity verification, real-time scoring, and getting things shipped. The previous administration's priorities were equity, bias auditing, and stakeholder process. Pitches that worked 18 months ago do not work now. Pitches that work now will need a different framing in 30 months. Do not assume any single policy stance will hold across cycles.

Closing skepticism, the legal exposure, and the next 18 months

Plenty of valid criticism exists. Critics including Public Citizen have pointed out that the administration's anti-fraud efforts have disproportionately targeted Democratic states, sometimes with shoddy data work. The administration acknowledged to the AP that a major data error went into the NY Medicaid fraud investigation. Generative AI models hallucinate, and an audit finding generated by an LLM that cites a deficiency that does not exist is not just embarrassing, it is potentially actionable under APA arbitrary-and-capricious standards.

The litigation pipeline is loading up. The Pieces Technologies Texas AG settlement set the template for AI vendors getting sued over performance representations. Lokken v. UHC at the District of Minnesota is currently working through aggressive discovery on UnitedHealthcare's AI claims tools and will produce precedent on what plaintiffs can demand from payer AI systems. The DOJ-HHS False Claims Act Working Group that re-launched in 2025 has signaled it will focus on AI-enabled billing and risk adjustment irregularities. Expect FCA actions specifically targeting providers and vendors who used AI tools that materially inflated reimbursements. Expect counter-suits from providers arguing that payer or government AI tools produced biased or erroneous denials. Expect at least one major FOIA fight over the training data and prompts used in the HHS single audit ingestion.

The next 18 months should produce, on the policy side, a CRUSH proposed rule probably in Q4 2026 or Q1 2027, a first round of CRUSH-related task orders against existing payment integrity vehicles, an OIG report critical of the May 21 announcement, and at least one Inspector General audit of the audit tool itself, which is the kind of recursive moment HHS specializes in. On the market side, expect

Codoxo to push more aggressively into government work, Cotiviti to extend its RAC role, two to four provider-side audit defense Series A rounds, and probably one acquisition of a small AI startup by a payment integrity incumbent at a frothy multiple.

For builders, the real opportunity is not chasing the headline. The headline is just permission. The opportunity is in the inverse of what the government is doing: building tooling that helps providers, plans, and grantees produce, defend, and verify their own audit position before the government's AI even sees it. That category has no clear leader. It is technically tractable. It has clear willingness to pay. It is sitting there waiting for the right founder to pick it up. The next 18 months will determine which two or three teams end up owning it.

Thoughts on Healthcare Markets &
Technology is a reader-supported
publication. To receive new posts and
support my work, consider becoming a
free or paid subscriber.

[← Previous](#)

Discussion about this post

[Comments](#)

[Restacks](#)



Write a comment...