DIAS™ EXECUTIVE RESEARCH SERIES · STRATEGY & FRAMEWORK WHITE PAPER

The Interoperability Tariff™

A cost-aware framework for EHR-integrated AI
across providers, payers, and pharma

The thesis in one line — *every time data crosses a boundary (out of the EHR, between platforms, and back into the chart) it pays a tariff, charged again at every border: site × platform × workflow.*

Philip Morales
Founder & Chief Transformation Officer
Digital Intelligence Advisory Services (DIAS™)

Version 1.3 · July 2026 · For Public Distribution

CONTENTS

Executive Summary

Two transitions are reshaping healthcare technology at once. First, electronic health records (EHR/EMR) are migrating from on-premise installations to the cloud. Second, AI solutions are being layered on top of those records to deliver administrative and clinical value — workflow dashboards, data integration, analytics, care-note generation, diagnostic recommendations, and medication-interaction screening.

Both shifts move cost from predictable, capitalized licenses toward variable, usage-based bills. The trouble is that leaders budget for the visible layer — model and token cost — while the real, recurring expense accumulates in the **interoperability tariff**: the toll paid every time data crosses a boundary — extracting it from the EHR, normalizing it, running inference, and, most expensively, writing results back into the chart. That toll is charged again at every site, every EHR platform, and every workflow.

This paper argues that the token layer is the *deflating* cost (Gartner projects roughly 90% lower inference cost by 2030), while the interoperability tariff is the *durable, compounding* cost, because it is driven by labor, vendor certification, and per-instance integration rather than by chip economics. The strategic response is a hybrid architecture: keep the high-frequency, PHI-heavy loop on edge / small language models inside the data boundary so it clears fewer borders, and reserve the cloud for de-identified, low-volume, heavy-lift reasoning.

The framework applies, with different emphasis, to all three healthcare buyers — providers, payers, and pharma — and the paper closes with a prioritization model, a 90-day roadmap, and a cost-driver checklist.

1. The Shift: Cloud EHR, with AI on Top

The first wave of enterprise AI was simple to procure: connect to a large cloud model, pay per token, and move on. In healthcare, that wave is arriving at the same time as a structural move of the EHR itself to the cloud. Tufts Medicine, for example, moved Epic and 42 integrated applications to AWS; AWS reports roughly four million patient records ingested in a 72-hour window. Oracle Health operates entirely on cloud infrastructure, and MEDITECH Expanse runs across AWS, Azure, and GCP.

Layered on top of the cloud EHR is an AI value layer that healthcare organizations increasingly treat as essential. The spend is real and concentrated: healthcare captured nearly half of all vertical AI spend — roughly $1.5B — and about 85% of healthcare generative-AI spend flowed to startups rather than incumbents (Menlo Ventures). Ambient clinical documentation alone represented an estimated $600 million in 2025 spend.

The problem is that both transitions quietly convert fixed costs into usage-based costs — and the largest usage-based line items are the ones that never appeared in the original business case.

The AI Value Layer on the EHR

In practice, the AI layer is asked to perform a recurring set of administrative and clinical jobs, each of which both reads from and writes to the record:

- **Administrative & operational dashboards** — throughput, scheduling, denials, and revenue-cycle views built on integrated EHR data.
- **Data integration & analytics** — unifying clinical, claims, and operational data for population and performance analytics.
- **Care-note generation** — ambient capture and structured note drafting returned to the chart for clinician review.
- **Diagnostic recommendations** — decision support surfaced in the clinician's workflow at the point of care.
- **Medication-interaction screening** — real-time checks that must be both fast and clinically validated.

Every one of these jobs depends on getting data out of the EHR and, critically, pushing results back in. That round trip is where the cost framework in this paper lives.

2. The Core Thesis: Every Crossing Pays a Tariff

It is tempting to model AI cost as the price of inference. But inference is the one component on a steep deflationary curve. Gartner projects that performing inference on a one-trillion-parameter model will cost providers more than 90% less by 2030 than in 2025, and that models will become up to 100 times more cost-efficient than the earliest versions — driven partly by inference-specialized silicon and edge devices.

The integration layer follows the opposite curve. It is built from interface engineering, terminology mapping, vendor certification, clinical validation, and ongoing maintenance — all labor- and license-driven inputs that do not benefit from semiconductor economics. Like a customs duty, the tariff is levied again at every border the data crosses, so as an organization adds sites, platforms, and workflows, this cost compounds rather than deflates.

The strategic implication: optimizing the model is necessary but insufficient. The leverage is in reducing the *crossings* — how much of the high-frequency loop you can keep local, standardized, and reusable, so data clears fewer borders and pays the tariff fewer times.

3. The Cost Framework: The Loop, Multiplied

Every workflow in the AI value layer runs the same four-step loop against the EHR:

1. **Extract** — pull data via batch (bulk FHIR $export) and transactional APIs or HL7 v2 messages.
2. **Normalize** — map to standard terminologies (LOINC, SNOMED CT, RxNorm, ICD-10) and reconcile data-quality differences.
3. **Infer** — run the model (edge / SLM or cloud) to produce the note, recommendation, or analytic output.
4. **Write back** — push the validated result into the chart, where it enters clinical workflow.

Each step is a border crossing, and each crossing carries a tariff. Reading is comparatively cheap; writing back is comparatively expensive. And then the whole loop is multiplied:

Operational Sites × EHR / EMR Platforms × Workflows = Total Integration Surface

Each new site and each new EHR platform is its own integration, certification, and maintenance obligation. Two hospitals can both run Epic, both certified for FHIR R4, and still return the same patient data in different shapes — so even "same-vendor" consolidation does not eliminate the normalization burden. Vendor certification typically runs $5,000–$15,000 and four to eight weeks per platform before a single inference is billed.

Cost is not the only thing that compounds. When AI agents run this loop autonomously, errors compound at every crossing — five 95%-reliable steps chain to roughly 77% end-to-end ($0.95^5 \approx 0.774$) — and a wrong write-back carries a far larger blast radius than a wrong read. That makes each crossing a governance decision as much as a cost one; the companion paper, *When the Loop Runs Itself* (Paper No. 2), introduces the Write-Back Autonomy Ladder™, which grants autonomy by tier, weighted to each crossing's blast radius.

The Total Cost of Ownership (TCO) Stack

The table below separates the one layer that deflates from the layers that endure. Most planning attention goes to the first row; most of the lifetime cost lives in the rest.

Cost Layer	What Drives It	Trajectory
Model / token inference	Chip & model efficiency	DEFLATES
Data egress from cloud EHR	Per-volume cloud charges	DURABLE
Interface / API & certification	$5K–$15K + weeks per platform	DURABLE
Terminology normalization & data quality	Ongoing mapping & remediation labor	DURABLE
Write-back validation & human-in-the-loop	Clinical review labor	DURABLE
Interface maintenance / version drift	Breaks on each EHR upgrade	DURABLE
Governance, audit, BAA & storage	Continuous compliance overhead	DURABLE

4. Why Write-Back Is the Hard, Expensive Part

The single most underestimated line item is write-back. Reading data from an EHR is well-trodden; writing data back requires validation logic, error handling, conflict resolution, and careful adherence to each EHR's business rules. Three forces make it expensive:

- **Technical asymmetry.** Standard FHIR support is strongest for reads; many write operations rely on transactional HL7 v2, proprietary APIs, or constrained documentation endpoints. Note that ~95% of US healthcare organizations still use HL7 v2 internally, so write-back rarely escapes legacy interfaces.
- **Clinical-safety burden.** Writing diagnostic recommendations or medication-interaction alerts into the chart engages clinical decision support, alert-fatigue management, liability, and potential Software-as-a-Medical-Device scrutiny. Human-in-the-loop review is effectively mandatory, which is recurring labor, not a one-time build.
- **Per-instance certification.** Each EHR vendor's marketplace (Epic Showroom / Connection Hub, Oracle Health Code Console, athenahealth) imposes its own review, cost, and timeline — and Epic's approval is per customer instance.

In short: the extract side scales with data volume; the write-back side scales with workflows, sites, and clinical governance — which is why it dominates the durable cost.

5. The Architecture Response: Keep PHI at the Source

If the durable cost is the loop, the design goal is to make the loop local, standardized, and reusable. The recommended pattern is a hybrid architecture organized around a PHI boundary.

What Runs Where

Zone	What It Does	What Runs Here
Inside the PHI boundary (in-VPC / on-prem)	The high-frequency read → normalize → infer → write-back loop	Interface engine; terminology normalization; edge / SLM inference; governance & audit
Cloud frontier (outside boundary)	Low-frequency, heavy-lift reasoning	Frontier LLMs; population analytics; research; training — on de-identified data only

Three benefits follow directly from keeping the loop inside the boundary:

- **Query data where it lives.** When patient data never leaves your infrastructure, third-party exposure shrinks — important given that 80%+ of stolen PHI is taken from vendors and software services rather than hospitals directly, and healthcare breaches average $7.42 million, the highest of any industry for 14 consecutive years (IBM, 2025).
- **Cut the token and egress bill.** On-prem / in-VPC small language models handle routine, high-volume tasks without per-token cloud fees or repeated data egress.
- **Real time at the point of care.** No server round trip means responsive ambient notes, coding, and interaction checks inside the clinician's workflow.

Design rule: the loop that touches PHI most often should stay local and cheap; the cloud should earn its cost only on de-identified, low-volume, heavy lifting. Build the integration as a *hub* (a shared interface and normalization layer) rather than point-to-point spokes, so the second and third EHR platform cost materially less than the first.

6. One Framework, Three Buyers

The same loop and the same boundary apply to providers, payers, and pharma — but the dominant cost driver differs by segment.

Segment	Where AI Works	Dominant Cost Driver
Providers	Ambient notes & coding written back to the chart; diagnostic / decision support; real-time medication-interaction screening	Write-back into Epic / Oracle Health / MEDITECH across many sites
Payers	FHIR prior-authorization APIs (CMS-0057-F); risk-adjustment & utilization analytics; payer-to-payer and provider data exchange	Regulatory FHIR APIs and cross-plan data integration
Pharma	Real-world evidence from de-identified EHR data; pharmacovigilance & safety-signal detection; medical-affairs and commercial analytics	Data licensing, de-identification, and large-scale compute

For payers in particular, the regulatory clock is now a cost event: CMS-0057-F requires FHIR-based prior-authorization, provider-access, and payer-to-payer APIs across Medicare Advantage, Medicaid, CHIP, and Marketplace plans, with major API requirements generally due January 1, 2027 and USCDI v3 (via FHIR US Core) due January 1, 2026. The Da Vinci CRD, DTR, and PAS implementation guides are the recommended framing rather than mandated named standards. These mandates convert interoperability from a nice-to-have into a budgeted obligation.

7. Honest Counterpoints — and Why the Thesis Holds

"Native EHR AI will absorb this."

EHR vendors are embedding their own AI (for example, native ambient documentation and agent features). Where a single vendor's native AI fits the workflow, it can avoid the egress and integration tax entirely. But native AI is vendor-locked and single-platform. Multi-EHR health systems (common after M&A), best-of-breed clinical needs, and payer/pharma data flows that span organizations still require the independent integration layer this framework describes. The thesis narrows to single-vendor, single-workflow cases — it does not disappear.

"Token cost is falling, so cost will solve itself."

Token deflation is real, and it is exactly why it is the wrong thing to optimize for. The cost that determines program economics — interface engineering, normalization, write-back validation, certification, and maintenance — is labor- and license-driven and does not deflate. Falling token prices actually strengthen the argument to move high-volume inference local: the marginal model cost is shrinking, so the differentiator becomes who controls the integration layer and minimizes the tariff.

"Shadow AI is already filling the gap."

It is — and that is a risk, not a solution. 40% of healthcare professionals report encountering unauthorized AI tools and 17% report using them — 57% combined (Wolters Kluwer, January 2026); shadow AI adds roughly $670,000 to the cost of a breach and 63% of organizations still lack formal AI governance (IBM, 2025). A governed hybrid architecture is the sanctioned alternative that displaces shadow AI.

8. Prioritization Model and 90-Day Roadmap

How to Prioritize Each Workflow

Score every candidate workflow on five dimensions. The higher the score, the stronger the case to run it locally on edge / SLM inside the PHI boundary.

Dimension	Question	Pushes Toward
Data sensitivity	Does it touch PHI?	Edge
Volume	Is it high-frequency / high-volume?	Edge
Latency	Does it need real-time response in workflow?	Edge
Write-back	Does it push results into the chart?	Edge
Reasoning depth	Does it need frontier-level reasoning at low volume?	Cloud

A 90-Day Path

5. **Weeks 1–3 — Map & meter.** Inventory AI workflows and EHR touchpoints. Meter egress, API, and inference spend. Classify each workflow by sensitivity, latency, and read-vs-write.
6. **Weeks 4–8 — Pilot the loop.** Pick one or two high-volume, PHI-heavy workflows. Stand up the extract → normalize → infer → write-back loop on edge / SLM at a single site.
7. **Weeks 9–12 — Hub, not spokes.** Centralize on a hub-and-spoke interface layer so the second and third EHR cost less. Add governance, audit, and write-back validation.
8. **Ongoing — Scale & govern.** Expand site by site, formalize AI governance, and treat version drift as a standing budget line rather than a surprise.

9. Cost-Driver Checklist

Use this list to pressure-test any EHR-integrated AI business case before approval. Each item is a cost that is frequently omitted from initial estimates.

- ☐ Data egress fees from the cloud EHR for every extract
- ☐ API / interface licensing and per-vendor, per-instance certification ($5K–$15K + 4–8 weeks each)
- ☐ Terminology normalization and ongoing data-quality remediation (LOINC, SNOMED, RxNorm, ICD-10)
- ☐ Write-back development plus recurring human-in-the-loop clinical validation
- ☐ Interface maintenance and version-drift remediation on each EHR upgrade
- ☐ Per-site and per-platform multiplication of all of the above
- ☐ Data duplication / storage if EHR data is copied into a lake or warehouse
- ☐ Governance, audit, BAA, and security overhead for each new data flow
- ☐ Provisioned vs. peak inference capacity (idle cost)
- ☐ Ongoing model evaluation and clinical re-validation
- ☐ Vendor lock-in and switching costs in the integration layer

Conclusion

Cloud EHR migration and the AI value layer on top both shift cost from predictable licenses to usage-based bills — and the surprises hide in the integration, not the model. The durable, compounding cost is the tariff on every crossing: extract, normalize, and especially write-back — charged again at every site, platform, and workflow.

The strategic answer is not to choose between cloud and edge, but to place each workload where it is cheapest and safest: keep the PHI-heavy, high-frequency loop local on edge / small language models inside the boundary, and reserve the cloud for de-identified heavy lifting. Build a hub, meter the spend, pilot one loop, and govern from day one.

Treat AI as a workflow that reads from and writes to the EHR — and the cost question answers itself.

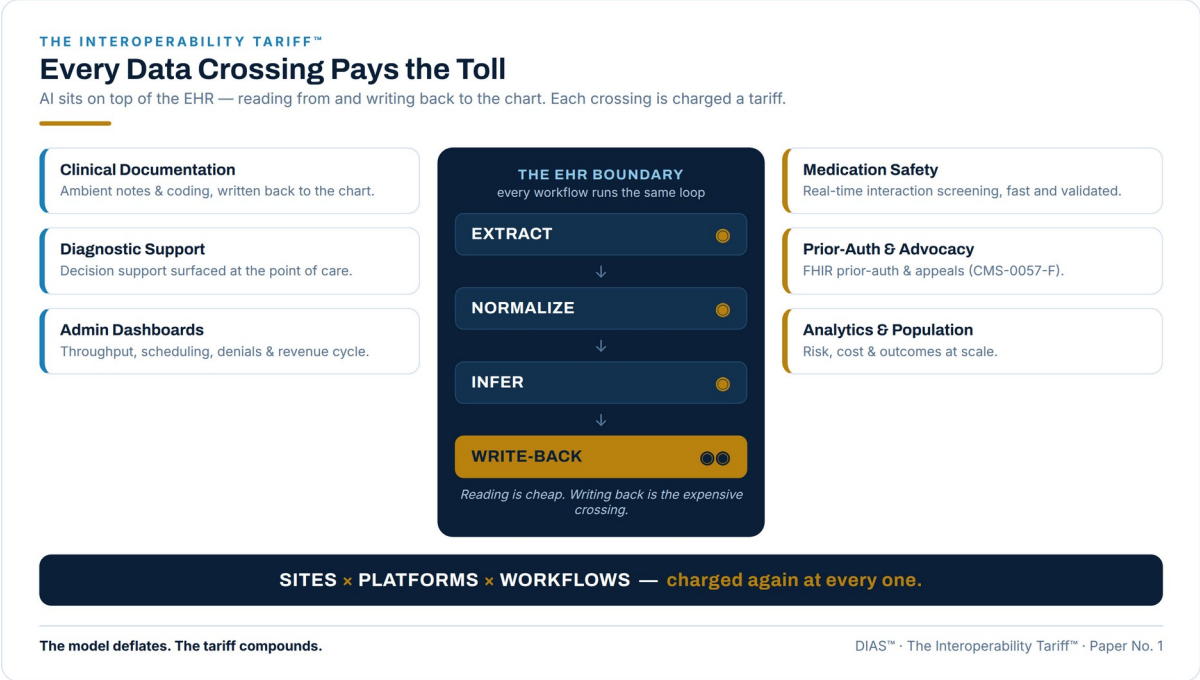

Figure 1 · The Interoperability Tariff™ — the recurring cost every data crossing levies across the healthcare AI stack.

References & Validation

Validation Statement

All regulatory claims were verified against primary government sources. Quantitative market, adoption, and implementation figures were verified against the best available 2025–2026 industry, peer-reviewed, or recognized secondary sources. Commercial market forecasts and vendor-reported implementation metrics are identified as directional estimates where appropriate.

Tiers — **Verified** (confirmed against the issuing organization's own publication), **As reported** (accurately attributed to the source that reported it), and **Directional** — are marked in each reference entry. **Of 22 sourced claims: 19 verified against a primary or issuing source; 3 accurately attributed as reported.**

Three claims were withdrawn in v1.3 rather than reported, because they could not be traced to an issuing source: hospital FHIR API use "56%→84%", "Epic per-instance approval 4–6 weeks", and "~2.2× faster healthcare AI adoption." Two claims were corrected for scope: the Menlo Ventures 85% figure is a share of healthcare generative-AI spend flowing to startups, not an adoption rate; and the ">80% of stolen PHI" figure derives from AHA / Chartis analysis of HHS-OCR breach data, not from SecurityScorecard.

Analytical contributions — the Interoperability Tariff™ framework, the integration-surface identity, the $0.95^5 \approx 0.774$ arithmetic, and the blast-radius framing — are original or independently verifiable and are excluded from the count. Market and integration figures are dated to 2025–2026 reporting and vary by environment; validate against your own vendor contracts and instance counts.

Sources

[Verified] IBM. *Cost of a Data Breach Report 2025*. Healthcare breach average $7.42M (highest 14 years; 279-day containment); shadow AI adds ~$670K; 63% of breached organizations either lack an AI governance policy or are still developing one.

[Verified] American Hospital Association, *2025 Cybersecurity Year in Review*; Chartis (2026), analyzing HHS-OCR breach data. More than 80% of stolen PHI records were taken not from hospitals but from third-party vendors, software services, business associates, and non-hospital providers and health plans; more than 90% of stolen PHI was taken from outside an EHR system. (SecurityScorecard's 2025 Global Third-Party Breach Report is a separate finding: healthcare recorded the most third-party breaches of any sector, at a 32.2% rate.)

[Verified] Menlo Ventures. *2025: The State of Generative AI in the Enterprise*. Healthcare captured nearly half of vertical AI spend (~$1.5B); ambient documentation ~$600M; ~85% of healthcare generative-AI spend flowed to startups rather than incumbents. (The 85% figure is a share of spend, not an adoption rate.)

[Verified] Wolters Kluwer (January 2026). 40% of healthcare professionals encountered unauthorized AI tools and 17% used them; 57% encountered or used shadow AI in combination.

[As reported] Taction (2026). ~95% of hospitals still use HL7 v2 internally; app certification typically $5K–$15K and 4–8 weeks per platform. CapMinds (2026). Read-only integration is materially simpler than write-back.

[Verified] CMS. *Interoperability and Prior Authorization Final Rule (CMS-0057-F)*, 89 FR 8758, February 8, 2024 (FR Doc. 2024-00895); effective April 8, 2024. Operational prior-authorization provisions generally due January 1, 2026; major API requirements — Provider Access, Payer-to-Payer, Prior Authorization, and Patient Access enhancements — generally due January 1, 2027. ONC HTI-1: USCDI v3 becomes the required Certification Program baseline January 1, 2026. Da Vinci CRD/DTR/PAS are implementation guides, not mandated named standards.

[Verified] Gartner (March 25, 2026). Inference on a one-trillion-parameter LLM projected to cost GenAI providers over 90% less in 2030 than in 2025; LLMs up to 100× more cost-efficient than comparable 2022 models. Gartner further

notes token-cost deflation will not pass through fully to enterprise customers, and that as token consumption rises faster than token costs fall, overall inference costs are expected to increase.

[Verified] Cloud migration reference: Tufts Medicine moved Epic and 42 integrated applications to AWS; AWS reports ~4M patient records ingested in a 72-hour window (Healthcare IT News reports 3M+ health accounts in 71 hours; figures differ by source and unit of measure). MEDITECH Expanse runs on AWS/Azure/GCP; Oracle Health Ignite APIs are cloud-native (FHIR R4). [Vendor documentation.]

[Analytical, excluded from the count] Multi-agent reliability & blast radius (2026). Chained-agent reliability compounds multiplicatively ($0.95^5 \approx 0.774$; independently verifiable). "Blast radius" is a standard site-reliability/security term applied here to interoperability crossings. See Paper No. 2, *When the Loop Runs Itself*, which introduces the Write-Back Autonomy Ladder™.