

CUSTOMER SUCCESS STORY

Clinical AI at scale, with PHI that never leaves.

Imaging, drug discovery, clinical documentation, and RAG over the medical record — four programmes, one non-negotiable constraint. Here is how a health system with a research arm put all four into production on **UPC AI Factory**.

62–67%

Healthcare AI adoption reached roughly 62–67% in 2026 — the fastest acceleration of any industry, up from about 38% in 2024. The question has shifted from whether to deploy clinical AI to where it is allowed to run.

BUILT ON UPC AI FACTORY

THE SITUATION

Four programmes, one immovable constraint

Momentum was not the problem: imaging models were reading studies, discovery pipelines were screening candidates, and clinicians were asking for ambient documentation. Every programme met the same wall:

- **PHI could not leave the boundary.** Prompts, embeddings, and weights derived from patient records fell inside HIPAA obligations and BAA scope.
- **Imaging data was too heavy to move.** Shuttling studies to remote GPUs added cost and hours to every training run.
- **Discovery jobs were bursty and costly.** On-demand GPU pricing made screening campaigns hard to fund and forecast.
- **Latency at the point of care.** Documentation and retrieval had to answer inside the visit, not after it.

THE DECISION

Bring the compute to the data

Rather than negotiate exceptions to move patient data outward, the system moved the GPUs inward — operated by UnitedLayer, governed by the health system.

“We were never going to win an argument about moving patient data. So we stopped having it, and moved the compute to where the data already lived.”

— CHIEF INFORMATION & DIGITAL OFFICER

CUSTOMER PROFILE

INDUSTRY

Healthcare & life sciences

PROFILE

Health system with a research & discovery arm

AI WORKLOADS

Imaging · Drug discovery · Clinical documentation · Record RAG

FOOTPRINT

Two primary data centres plus a DR site

TENANCY

Isolated vDCs — Clinical, Research, Imaging

TIME TO PRODUCTION

12 weeks, end to end

GOVERNANCE

PHI in-boundary, BAA scope, auditable lineage

SOLUTION

UPC AI Factory, fully managed

WHAT WAS DEPLOYED

One governed estate, clinical and research side by side

AI services

OpenAI-compatible Router, Model Registry, managed Vector DB for record retrieval, and GPU Workbench for research teams — private models behind familiar interfaces.

Compute

Dedicated **GPU pools with MIG partitioning**, separating bursty discovery training from the steady inference that clinical documentation depends on.

Data

High-performance storage for imaging archives, model artefacts, and vector indexes, **co-located with GPU** so studies are read where they are stored.

Governance & operations

Backup, DR, and a **cyber vault** for clinical models; observability and AIOps into existing tooling, with **FinOps dashboards** reporting cost per million tokens by programme.

THE RESULTS · YEAR ONE

↓ **42%**

Lower TCO for AI workloads

Capacity redirected into more research, not more invoices

↑ **40%**

Faster model training

Single-tenant, RDMA-native GPU fabric (35–40%)

<100ms

p95 inference latency

Documentation and retrieval inside the encounter

100%

PHI, models, and embeddings inside the tenant boundary

No external exposure — evidence a compliance officer and an external auditor can both inspect in one place.

~**60%**

Less platform toil

One operated stack replaced the tooling clinical IT and research computing had been maintaining separately.

Fixed

Cost per million tokens, by programme

Imaging, discovery, documentation, and retrieval each carry a known cost, making grant and budget planning realistic.

WHY THIS MATTERS IN HEALTHCARE & PHARMA

In a regulated clinical setting, where a model runs is part of the control environment. Keeping the estate in-boundary turns AI governance into something demonstrable rather than asserted.

PHI kept in-boundary

HIPAA & BAA scope

Model lineage & versioning

Auditable access control

Research data segregation

GxP-aligned change control

WHERE THE SYSTEM IS TAKING IT NEXT



Multimodal imaging

Pairing radiology models with report text and history, which only works when both sit inside the same boundary.



Discovery at larger scale

Fixed GPU capacity turns screening campaign size into a scientific decision rather than a budget approval.



Documentation across specialties

Extending ambient capture beyond the first service lines, with latency that holds as concurrency grows.

BOTTOM LINE

UPC AI Factory let clinical and research AI scale together — inside the boundary, on a cost base the system controls.

Ready for enterprise AI without compromise?

Book a walkthrough, or ask us to scope a private AI factory for your workloads.

UPC AI FACTORY



EMAIL

contact@unitedlayer.com



PHONE

+1-888-853-7733

[UNITEDLAYER.COM](https://unitedlayer.com)