

CUSTOMER SUCCESS STORY

A SaaS platform made AI a feature, not a cost centre.

Their AI copilot and semantic search shipped fast and got used hard — then inference cost started eating gross margin and enterprise deals stalled in security review. Here is how they moved the AI tier onto **UPC AI Factory** and turned it into a margin story.

BUILT ON UPC AI FACTORY

THE SITUATION

The AI features worked. The economics did not.

Adoption was the good news: the in-product copilot and semantic search were among the most-used features in the platform, and AI had become the headline in every renewal conversation. The problems arrived with the scale:

- **Margin erosion.** Token and GPU spend grew faster than seats, pulling gross margin down quarter over quarter with no reliable way to forecast it.
- **Deals stuck in security review.** Enterprise buyers wanted written assurance that their data was isolated and never used for training — hard to give on shared infrastructure.
- **Latency inside the product.** Copilot responses sat behind a spinner long enough to be noticed, and shared GPUs made p95 unpredictable at peak.
- **An ops burden on a small team.** Model serving, vector search, GPU scheduling, and observability were maintained by engineers who were meant to be shipping features.

THE DECISION

Own the inference layer, sell the isolation

Instead of re-pricing the product around a volatile cost base, they moved inference onto dedicated infrastructure — and turned the isolation into something the sales team could sell.

“The moment we could put our inference on dedicated infrastructure, two things changed: our margin stopped moving around, and our security questionnaire stopped being an obstacle.”

— VP OF ENGINEERING

WHAT WAS DEPLOYED

An inference tier they control, operated for them

AI services	OpenAI-compatible Router, Model Registry, managed Vector DB, and GPU Workbench — so the existing application pointed at private models with a config change, not a rewrite.
Compute	Dedicated GPU pools with MIG partitioning, splitting steady inference from bursty fine-tuning so a training job could never slow the product.
Tenancy	Separate vDCs for the shared tier and the enterprise isolation tier, with regional placement for customers holding EU data-residency clauses.
Resilience & operations	Backup, DR, and a cyber vault for model artefacts; observability and AIOps into existing tooling, plus FinOps dashboards reporting cost per million tokens by feature.

CUSTOMER PROFILE

INDUSTRY

B2B SaaS, multi-tenant platform

PROFILE

Enterprise software vendor, multi-region customer base

AI WORKLOADS

In-product copilot · Semantic search · Document summarisation

FOOTPRINT

Two primary regions plus a DR site

TENANCY

Isolated vDCs for shared tier and enterprise tier

TIME TO PRODUCTION

12 weeks, end to end

COMMERCIAL MODEL

Single license, fixed monthly charge

SOLUTION

UPC AI Factory, fully managed

THE RESULTS · YEAR ONE

↓ **42%**

Lower TCO for AI workloads

Cost that flows straight into gross margin

↑ **40%**

Faster model training

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

<100ms

p95 inference latency

Copilot and search that feel instant in-product

100%

Customer data and models inside the tenant boundary

No external exposure — the security questionnaire became a differentiator instead of a delay.

~60%

Less platform toil

One operated stack handed the AI infrastructure back to UnitedLayer and the roadmap back to the engineers.

Fixed

Cost per million tokens, by feature

AI spend became a line in unit economics rather than a variable the finance team had to explain each month.

WHY THIS MATTERS FOR A SAAS BUSINESS

For a software vendor, private inference is not only an infrastructure decision — it changes what can be written into a contract, and what the AI tier does to the P&L.

No training on customer data

Per-tenant isolation

Regional data residency

Predictable COGS per seat

Enterprise tier as a paid SKU

SOC 2 · ISO 27001 · GDPR aligned

WHERE THE PLATFORM IS TAKING IT NEXT



A private AI tier as a SKU

Packaging dedicated inference as a premium plan for regulated customers who will pay for isolation in writing.



Regional expansion

Standing up additional regions to meet residency clauses without renegotiating the cost base each time.



Agentic workflows

Longer-running, multi-step features that only make sense economically on fixed-cost GPU capacity.

BOTTOM LINE

UPC AI Factory turned a volatile AI cost centre into a predictable, isolated, enterprise-ready part of the product.

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)