

FINANCIAL SERVICES

ILLUSTRATIVE SCENARIO

Frontier models beside the trading floor

A trading firm wanted large models to review contracts and help its support teams. Client documents and trading data could not touch a public API, and the bill could not grow with every new use.

ORGANIZATION

Trading and advisory firm

SITES

Trading site and branches

FIRST WORKLOADS

Contract review, support copilot

MODELS

Frontier LLMs

The situation

Legal and operations teams were reviewing contracts line by line, and branch support staff searched policy documents by hand. Early tests with public AI services showed the value, but compliance flagged the contractual risk of sending client documents off site, and finance flagged the per-token bill that would grow with every team that joined.

The turning point

The firm set two conditions for any AI rollout: sovereignty it could prove to auditors, and a cost it could budget. Usage was going to grow quickly once people saw it work, and the firm wanted that growth to add value, not cost.

What was at stake

Data boundary

Client documents could not leave the firm.

Audit

Every action had to be traceable.

Cost

Per-token pricing would grow with adoption.

Illustrative scenario based on typical UnitedEdge® AI engagements. Not a specific customer.

HOW IT UNFOLDED

UPC deployed a dual-engine frontier model at the trading site and a smaller copilot at the branches, all under one tenant policy.

01

Workshop

Picked contract review and a branch support copilot, with policy by model origin.

02

Dual-engine frontier

The largest model ran across both AI Engines at the trading site for contract review.

03

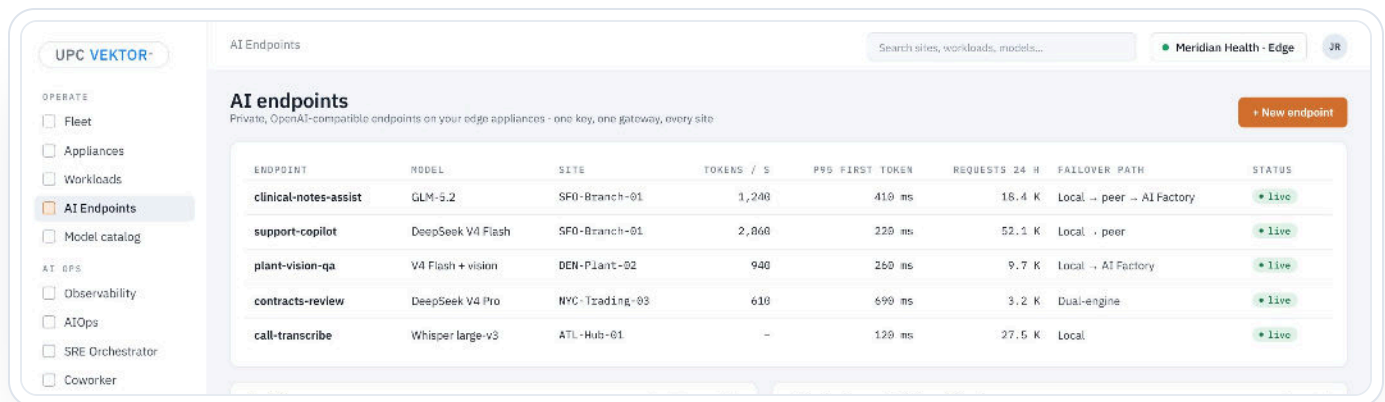
One URL for developers

Teams pointed existing OpenAI-style code at the private endpoint and kept their code.

04

Prove it monthly

The value report compared usage with public API list prices each month, under one fixed subscription.



ENDPOINT	MODEL	SITE	TOKENS / \$	P95 FIRST TOKEN	REQUESTS 24 H	FAILOVER PATH	STATUS
clinical-notes-assist	GLM-5.2	SFO-Branch-01	1,240	410 ms	18.4 K	Local → peer → AI Factory	• live
support-copilot	DeepSeek V4 Flash	SFO-Branch-01	2,860	220 ms	52.1 K	Local → peer	• live
plant-vision-qa	V4 Flash + vision	DEN-Plant-02	940	260 ms	9.7 K	Local → AI Factory	• live
contracts-review	DeepSeek V4 Pro	NYC-Trading-03	610	690 ms	3.2 K	Dual-engine	• live
call-transcribe	Whisper large-v3	ATL-Hub-01	-	120 ms	27.5 K	Local	• live

Private, OpenAI-compatible endpoints at each site, with failover paths and live latency.

WHAT CHANGED

AREA	BEFORE	AFTER
Data boundary	Contractual risk with APIs	→ Sovereign mode enforced on site
Model choice	One vendor's models	→ Curated open models, policy by origin
Cost	Per-token, grows with use	→ Fixed monthly, unmetered tokens
Audit	Scattered logs	→ Every action in one audit log

WHAT'S NEXT

The firm is extending the copilot to every branch and building approval-gated agents for routine operations reports.

Illustrative scenario based on typical UnitedEdge® AI engagements. Not a specific customer.