neysa | innoviti

Engineering a **60%** TCO reduction in field service operations with custom multimodal AI

Innoviti deployed Qwen 3.0 VL on Neysa Velocis to automate service quality assurance across 50,000+ large format merchant stores, delivering 96% verification accuracy and sub 30-second response times.

NEYSA SETUP

Service
Velocis Managed Inference Engine

Model
Qwen 3.0 VL
Custom multimodal, fine-tuned

Scale
>7,000
Service ticket logs / day

OUTCOME

60%
Total cost of ownership reduction

Sub 30s
Report processing time

0
False positives in production

96%
Output accuracy across all field reports

OVERVIEW

Innoviti maintains thousands of PoS terminals across 50,000+ merchant stores. Every field visit produces a report covering whether the terminal got fixed, the agent showed up, the job sheet is stamped for billing, and the product is intact.

Manual verification became a severe bottleneck. An early attempt to automate on a general purpose cloud provider hit unpredictable latency and unsustainable token costs.

By partnering with Neysa to deploy a custom Qwen 3.0 VL model on Velocis, Innoviti now audits 7,000+ field reports daily at 96% accuracy and sub 30 second latency, cutting total cost of ownership by 60%.

hello@neysa.ai neysa.ai

CHALLENGE

> " You can not run a real time service operations setup on infrastructure you cannot see. General purpose cloud APIs kept us locked out of the serving stack, making it impossible to diagnose latency spikes, manage cost and truly control our model's behaviour. "

Girish Varadarajan
Chief Data and AI Officer, Innoviti

The workflow looked simple on paper: send a field report to a model, get a verification back. But scaling to thousands of image-heavy reports a day, with agents waiting at the store, exposed what general-purpose cloud APIs cannot deliver.

01. Control

Operating blindly in black-box infrastructure

Locked out of the serving stack, Innoviti could not access hardware logs to troubleshoot latency, finetune the model, or disable unnecessary thinking modes to optimise for speed.

02. Economics

Paying a scalability penalty

As daily audit volume grew, token consumption surged and cost per token skyrocketed, a structure that would destroy unit economics at scale.

03. Latency

Unpredictable latency on shared APIs

Field agents needed answers while still at the store, but shared cloud APIs fluctuated with network traffic, making sub 30-second response times impossible to guarantee.

The Field AI Trilemma — Every prior option forced a critical trade-off

	General cloudAPIs	Self-hosted shared cloud	Required for production
Stack visibility hardware & serving logs	• Black box	• Partial	• Full root access
Latency report turnaround	• Unpredictable	• Variable	• < 30s guaranteed
Unit economics cost per token at scale	• Unsustainable	• Predictable	• Predictable

hello@neysa.ai neysa.ai

TRANSFORMATION

> " Neysa provided the goldilocks zone we were looking for. They gave us the exact balance of root level control, dedicated compute, and hands-on engineering support to get our system ready for production."

Girish Varadarajan
Chief Data and AI Officer, Innoviti

Running a custom Qwen 3.0 VL model on Neysa infrastructure, Innoviti turns 7,000+ daily field reports into verified service records with 96% accuracy, closes tickets while agents are still at the merchant site, and keeps complete engineering control over the stack.

01. Control

White-box sovereign infrastructure

Root-level access to the serving stack lets Innoviti's engineers monitor memory in real time, troubleshoot hardware errors directly, and stabilise a fine-tuned LLM setup for the long term.

02. Stack

Co-engineered multimodal optimisation

The orchestration layer was tuned for Qwen 3.0 VL's large context windows, with proprietary data integrated so the model reliably recognises Innoviti's specific store signboards, product stickers, accessories, and stamped job sheets.

03. Economics

Predictable unit economics at scale

A flat, dedicated compute model holds every image-heavy inference under sub 30-second latency while cutting total cost of ownership by 60%, giving Innoviti the financial predictability to scale to 7,000+ daily automated audits.

MEASURED OUTCOMES

Production benchmarks on Velocis Managed Inference

TCO reduction

60%

Total cost of ownership vs general-purpose cloud approach.

Report turnaround

Sub 30s

End-to-end field report verification while agents are on-site.

Daily volume

> 7,000 tickets/day

Field service reports processed and verified automatically.

Accuracy

96% verification

Sustained across all image-heavy multimodal field reports.

False positives

0

False positives in production across all audited reports.

Merchant stores

50,000 +

Large-format stores covered across Innoviti's service network.

Start your production AI deployment with Neysa

Get in touch with our experts.

Request a demo | hello@neysa.ai

neysa