

From Network Silos to Intelligent Operations

Building AI Resilience with AGIOne



Context

Telecommunications operators sit at an unusual intersection. They are among the heaviest producers of operational data, including network telemetry, customer interactions, and billing and fraud signals, and increasingly among the heaviest consumers of AI to make sense of it. Mobile and fixed network operators now use AI for network fault prediction, capacity planning, customer care automation, fraud and revenue assurance, and multilingual support across markets.

A regional telecommunications group, operating mobile, broadband, and enterprise connectivity services across several countries, reached this exact inflection point. What began as a series of promising AI pilots within separate business units, network operations, customer experience, and enterprise sales, evolved into a fragmented AI estate that constrained both efficiency and resilience.

This reflects a broader shift across the industry. As AI adoption accelerates across network and customer facing functions alike, infrastructure fragmentation is becoming a limiting factor rather than an enabler.



Challenge

The group's initial approach mirrored a common pattern among large telecommunications organisations. Each business unit deployed AI independently: a customer service chatbot built on one closed source model for the consumer mobile brand, a separate model for network fault diagnosis, a further model fine tuned on enterprise contract data for the business sales team, and country specific deployments to satisfy local data residency requirements.

While each application functioned adequately in isolation, the fragmented approach introduced systemic risk:

- **Underutilised, siloed compute.** GPU and compute resources were siloed by business unit, with utilisation frequently below 25% in any single stack while other teams queued for capacity elsewhere in the same group.



- **Fragmented integration.** Each country operation maintained its own bespoke integration code to call external model providers, multiplying technical debt with every new market launch or regulatory change.
- **A single point of failure.** Dependency on a single external provider for customer facing AI created a critical point of failure. During a nationwide network upgrade weekend, when call volumes to customer service peaked, the primary model provider experienced a multi hour outage. IVR triage, chatbot deflection, and agent assist tools were affected at the same time, and call centre queue times rose sharply until the fault was resolved.
- **Inconsistent governance.** Compliance teams struggled to demonstrate consistent governance across markets, since customer data, call transcripts, and network telemetry were being routed to different models under different data handling terms depending on which team had built the tool.

Increasingly, the group was not constrained by the availability of capable models, but by the absence of a shared, governable layer to operate them reliably across markets and business units.

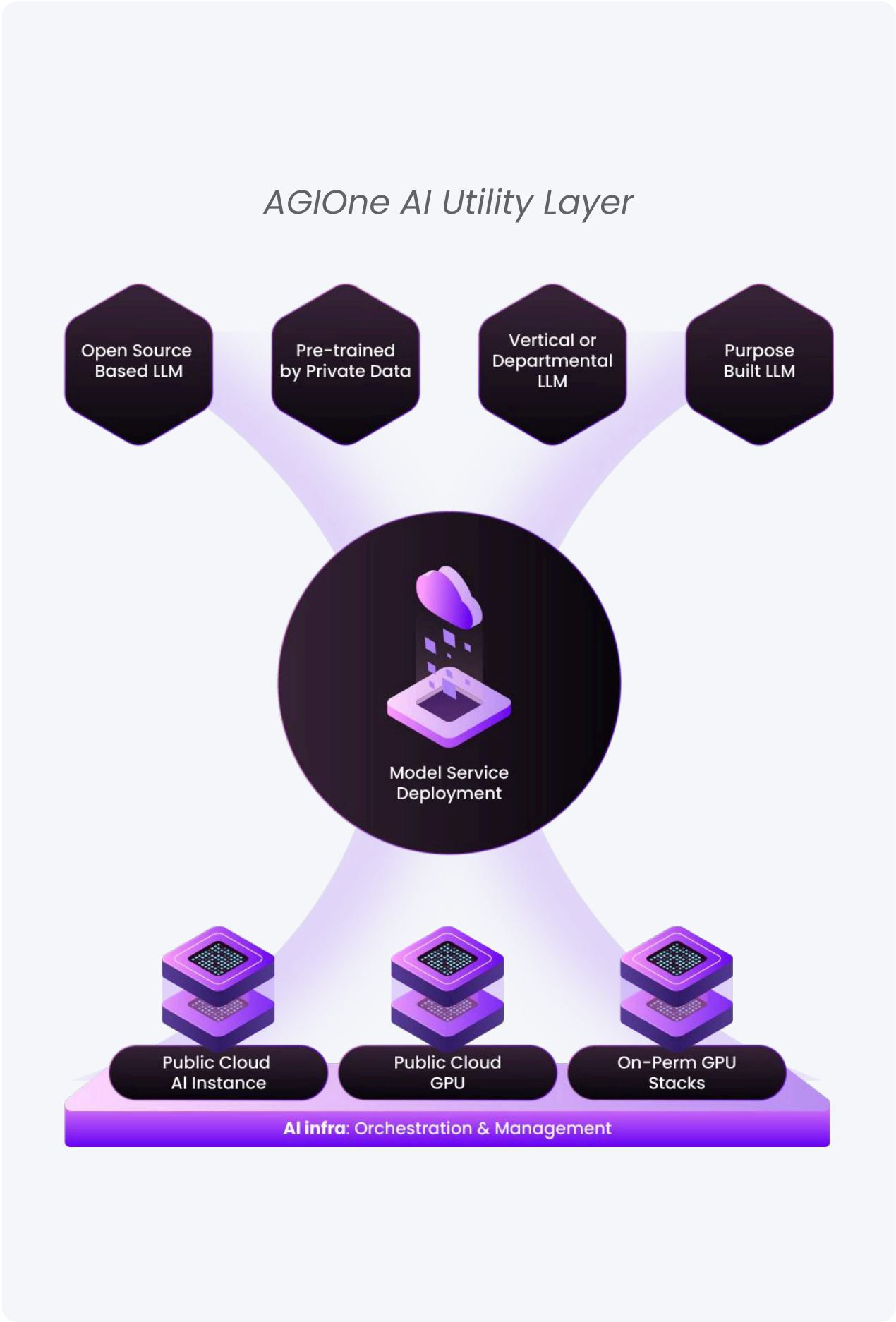
Approach: Establishing an AI Utility Layer with AGIOne

To address these limitations, the group implemented AGIOne as a unified AI gateway and orchestration platform spanning its markets. This marked a shift from application centric, per country AI stacks to a shared, policy driven operating model.

AGIOne introduced a shared GPU and model orchestration layer across the group's on premise data centres and multi cloud environments, turning previously siloed national infrastructure into a single elastic resource pool. Compute could be allocated dynamically to whichever workload needed it most, whether that was a surge in customer service volume in one market or a batch network diagnostics run in another.

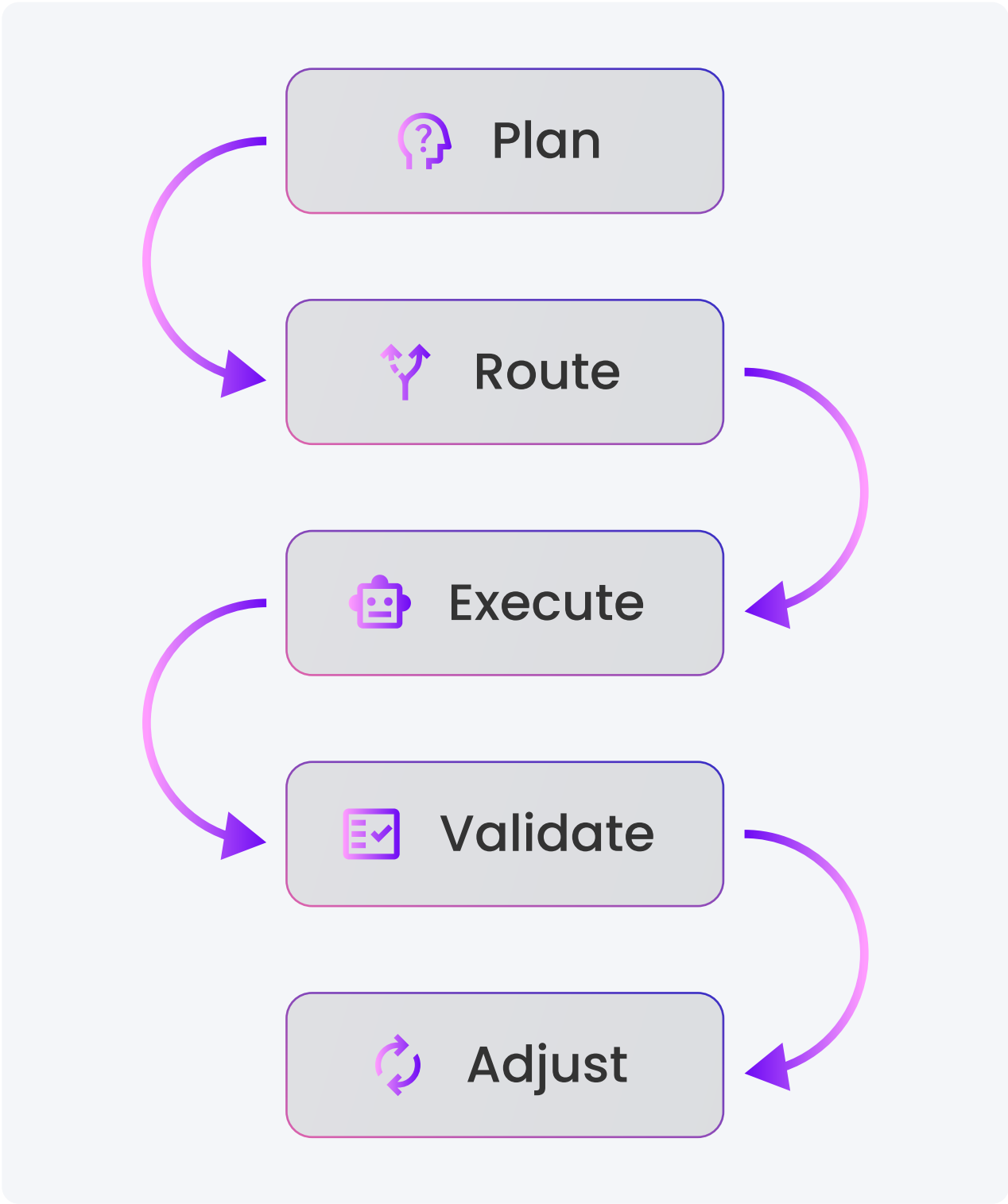
A unified model layer brought together open source models, privately hosted models trained on the group's own network and customer data, and models selected for specific regulatory environments, all behind one API. Intelligent routing selected the appropriate model for each request based on task type, latency requirements, data residency rules, and cost, without requiring engineering teams to hard code provider specific logic for every market.

Automated failover meant that when a primary model became degraded or unavailable, whether due to a provider outage, rate limiting, or scheduled maintenance, traffic was rerouted automatically to an alternative model. Customer facing workflows such as IVR triage, chat support, and agent assist continued operating without a visible interruption to customers.



The architecture also supported hybrid elasticity. When on premise GPU capacity in a given market reached its limit, workloads extended into public cloud GPU capacity without interrupting live operations, a capability particularly relevant during predictable demand peaks such as network upgrades, national holidays, or major sporting and cultural events that drive call and data volumes.

Operational Shift: From Deployment to Continuous AI Operations



A key transformation was the move from isolated AI deployments to full AI operations across both network and customer facing functions. Workloads were routed automatically based on latency requirements (real time network fault detection cannot tolerate the same latency budget as offline churn analysis), cost, and data sensitivity, with customer personal data and call recordings routed only to models meeting the relevant jurisdiction's data handling requirements.

Support for agentic workflows introduced a further level of capability. An AI agent could triage a customer reported network fault, correlate it against live network telemetry, check for known outages in the area, and either resolve the query directly or escalate it to a human engineer with full context attached, iterating through planning, action, validation, and adjustment rather than a single request and response exchange.

This signals a broader trend across the industry. As AI systems take on more autonomous roles in network operations and customer care, infrastructure must support not just execution, but coordination and adaptability across every market the group serves.

Outcome: From Cost Burden to Strategic Capability



Higher GPU Utilisation

Consolidated fragmented national and departmental capacity into shared pools, lifting utilisation and reducing duplicated hardware spend.



Faster Rollout Across Markets

Reduced the time needed to launch new AI powered customer service and network operations tools in each country.



Consistent Data Governance

Centralised routing and audit logging strengthened compliance confidence with regulators across every market served.



Resilience During Peak Demand

Customer facing and network operations workflows continued through provider outages and seasonal demand spikes.

Closing Insight:

From Connectivity Provider to AI Resilient Operator

This case reflects a critical inflection point in enterprise AI adoption within telecommunications. As operators increasingly compete on network quality, customer experience, and operational efficiency rather than on connectivity access alone, AI has become embedded in the systems that keep networks running and customers served.

Early success with isolated AI pilots is no longer sufficient. The ability to operate AI reliably across markets, business units, and regulatory regimes is becoming the defining differentiator.

By consolidating compute, unifying model access, and automating orchestration across its markets, the group positioned itself to support the next phase of its own transformation, one in which AI agents increasingly triage network faults, support customers around the clock, and inform network investment decisions, without tying that capability to any single external provider.

For telecommunications operators, the real goal is not ensuring every model call succeeds. The goal is ensuring the network and the customer stay connected when they do not.

ABOUT US

AGIOne & OnePro

AGIOne is OnePro Cloud's enterprise AI infrastructure product line, built to help organisations operate AI agents reliably across fragmented, multi-model environments at production scale.

