Cloud Native AI + Inference Day



Who controls your inference?

A field guide to sovereign AI

Bjorn Hovland, President, CIQ

**If your model went dark tonight,
who decides when it comes back?**

November 9, 2026

Sovereignty is a control architecture

Four independent decisions, made workload by workload

01	**Model**	Who holds the weights?
02	**Inference stack**	Who operates the runtime?
03	**Data path**	Where can the data travel?
04	**Update cadence**	Who approves production changes?

Who holds each key today?

©2026 CIQ INC.

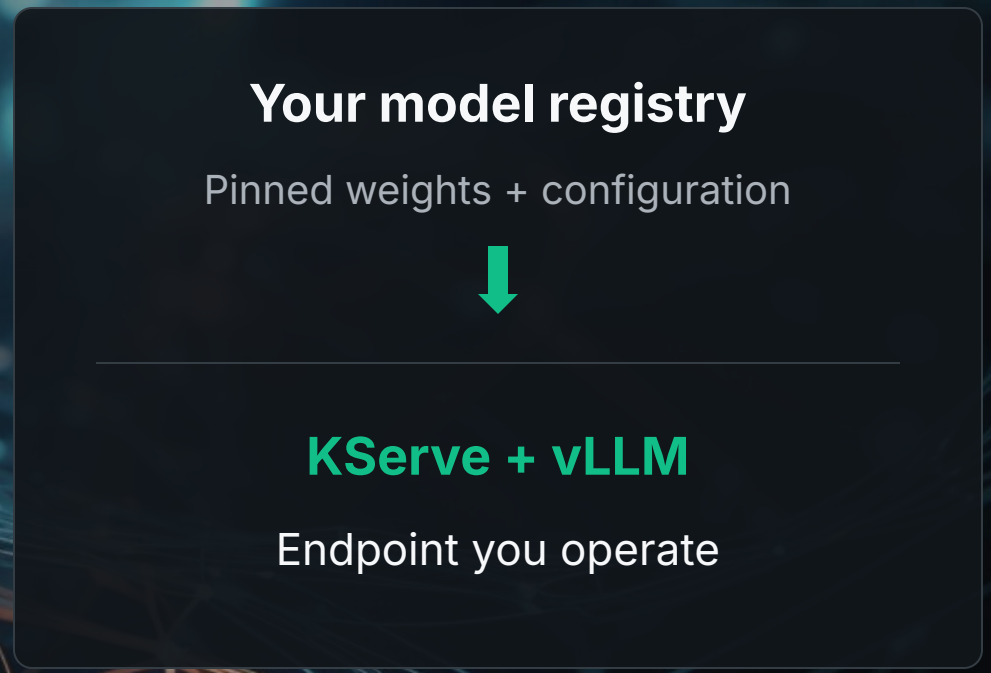
CIQ

Own the model

Weights you hold. A version you pin.

- **Keep the weights in your storage**
- **Pin the model and serving image**
- **Review the license and test quality**

Frontier APIs remain an option
where the workload needs them.



CIQ EXAMPLE Fuzzball makes the model a workflow parameter

©2026 CIQ INC.

CIQ

Own the inference stack

Control runs from the kernel to the serving endpoint

Serving + routing	vLLM, KServe, Gateway API Inference Extension
Kubernetes	GPU scheduling and placement
OS + GPU software	Kernel, drivers, CUDA, runtime
Hardware	Compute, network and storage

Your team decides

Deployment

Performance tuning

Upgrades and recovery

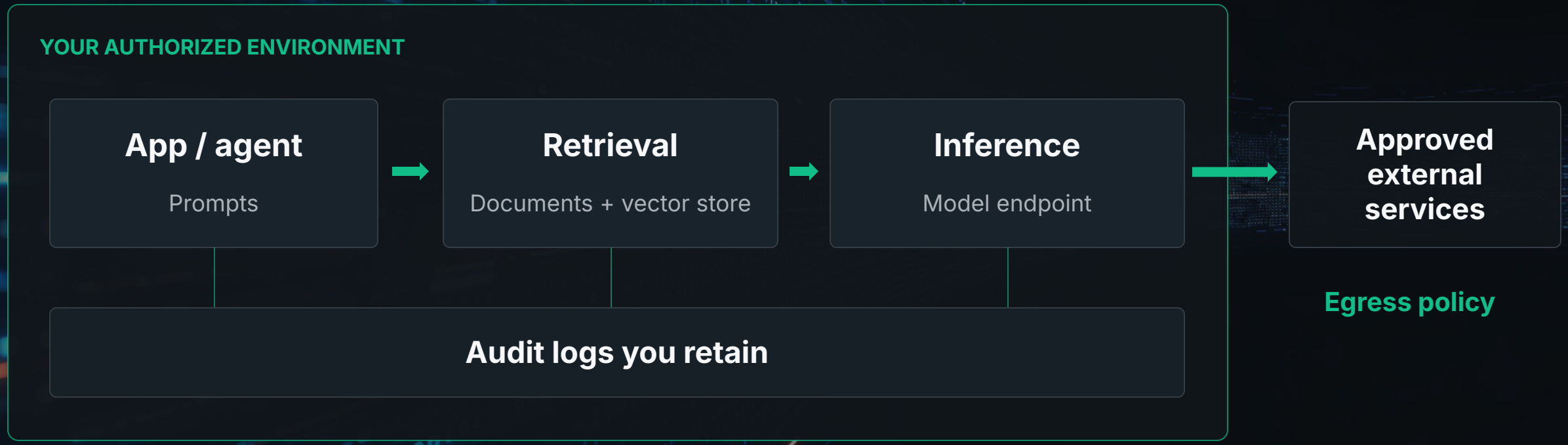
CIQ EXAMPLE RLC Pro AI for the OS and GPU stack. Stratum for Kubernetes.

©2026 CIQ INC.

CIQ

Own the data path

Know where prompts, retrieved documents and logs travel



Cilium network policy, egress controls and scoped identities

CIQ EXAMPLE Fuzzball runs inference and document retrieval inside your environment

©2026 CIQ INC.

CIQ

Own the update cadence

Production changes when your pipeline promotes it

GitOps, signed artifacts, SBOMs and build provenance



CIQ EXAMPLE SKE packages pinned components with signatures, SBOMs and provenance

©2026 CIQ INC.

CIQ

Control has operating costs

Capability, utilization and on-call ownership all matter

	Self-managed	Hosted API
Capability	Open-weight quality varies by task	Frontier models may lead on demanding tasks
Economics	GPU utilization and staffing drive total cost	Usage-based spend suits variable demand
Operations	You own capacity, patches and recovery	Provider runs the serving infrastructure

A mixed architecture is often the right outcome

©2026 CIQ INC.

CIQ

Four decisions. One workload

Hard requirements first. Then quality, cost and operations.

Model

Would losing this model for two weeks stop the business?

Inference stack

Do latency, throughput or GPU cost determine viability?

Data path

Must prompts or documents stay inside a defined boundary?

Update cadence

Would an unapproved change create audit, safety or accuracy risk?

Pick one workload this quarter.
Run it end to end under your control.

©2026 CIQ INC.

CIQ