



Sovereign AI Dinner

Private AI for Regulated Enterprises

By Marco Pavone, Kyle Watson
& Remco van der Horst



Il Menù

L'Aperitivo

L'Antipasto

Il Primo

Il Secondo

Il Formaggio e/o Il Dolce

Il Caffè con Il Digestivo

STARTER

Artisan bread with olive oil, balsamic vinegar, and smoked flakey sea salt

Shark Bay scallops with dashi beurre blanc and toasted quinoa

Four seasons wombok salad

MAINS

Red wine braised lamb shank with mashed potatoes

and walnut roasted heirloom carrots
artichoke and garlic yogurt

Seared duck breast à l'orange with mandarin supreme and parsnip purée

Stir fry black bean sake clams

Grilled black cod miso

DESSERT

Chocolate



Meet the Master Chefs



Marco Pavone

Senior Cloud Native Engineer



Kyle Watson

Cloud Native Engineer



Remco van der Horst

AI Strategy Consultant



No seat at the Table

Regulated enterprises deserve a seat at the table

Not every enterprise can cook in someone else's kitchen.

- Uncontrolled AI adoption is already happening
- Your data leaves the building
- Inflexible vendor lock-in



Data Sovereignty is a Hard Requirement

Absolute control over data, residency, and audit trails is a non-negotiable hard requirement.

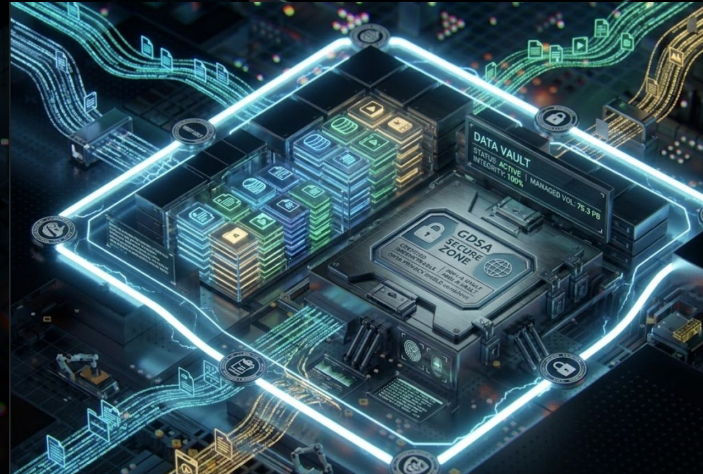
DORA

NIS2

BIO

GDPR

EU AI Act





The Kitchen

We like local suppliers, but serve to our clients taste





Extend instead of Reinvent

Focus on the dish instead of building a new kitchen



Build from scratch & integrate yourself

- High complexity & risk
- Expensive to maintain
- Time-consuming integration
- No guaranteed outcome



On top of platform

- Simple & fast deployment
- Proven reference architectures
- 100% Open-source foundation
- One contract.
Multitude of partners



Meet FinanceCo

Hungry for AI. Bound by regulation

A regulated enterprise that needs AI on its own terms.

FINANCECO

Regulated

DORA, GDPR and the AI Act apply

Sovereign

models and documents stay inside

Five departments

HR, Legal, Sales, Procurement, IT

TODAY'S MENU

Il Primo

answers from your own documents

Il Secondo

an agent that researches regulation





The Smart Retrieval system

Semantic natural language search with references

FinanceCo's staff want AI
Without a safe and smart option, they'll use public chatbots

Ask: "I like to chat with our company data"

Sources: HR, Legal, Sales, Procurement and IT Support documents

Your documents stay private. Zero data leaving the cluster.



Red Hat AI Accelerator

OpenShift AI, ready to serve. Every layer deployed from git

How the RAG system uses the platform

Your AI workload

The RAG system: Llama Stack, Llama 3.2 3B, Llama Guard 3 8B, PGVector

Red Hat OpenShift AI

KServe serves the models, with vLLM as the runtime

NFD + NVIDIA GPU Operator

Find the GPUs and install the drivers

Red Hat OpenShift

The container platform everything runs on

NVIDIA L40S GPUs

2 × 48GB: one for the LLM, one for the guardrail

ACROSS THE STACK

OpenShift GitOps

ArgoCD installs the operators and workload from git

User Workload Monitoring

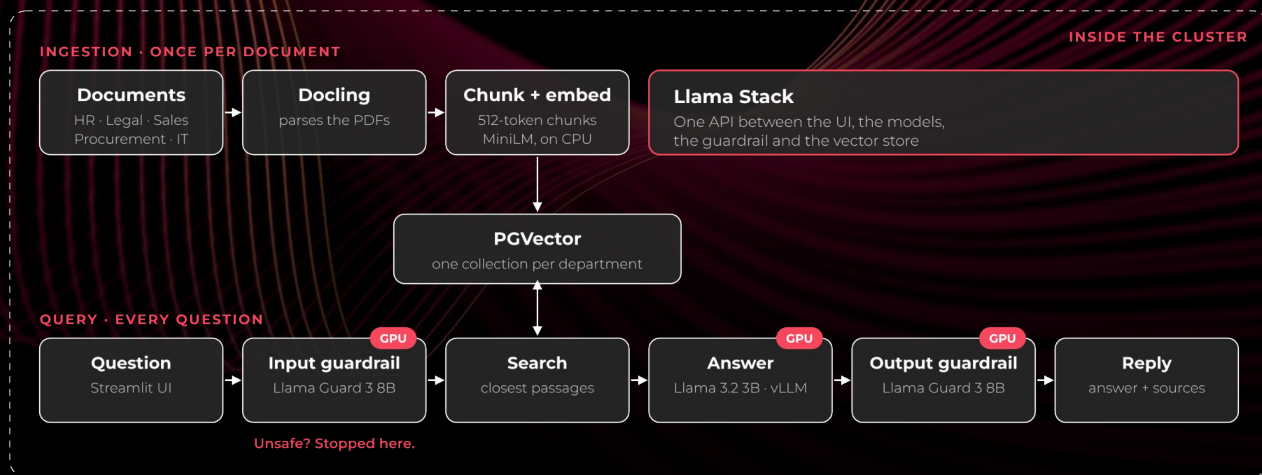
Prometheus metrics for your workloads



The RAG Architecture

Two pipelines: one ingests documents, one answers questions

Everything we deploy lives in git: reviewed, reproducible, auditable

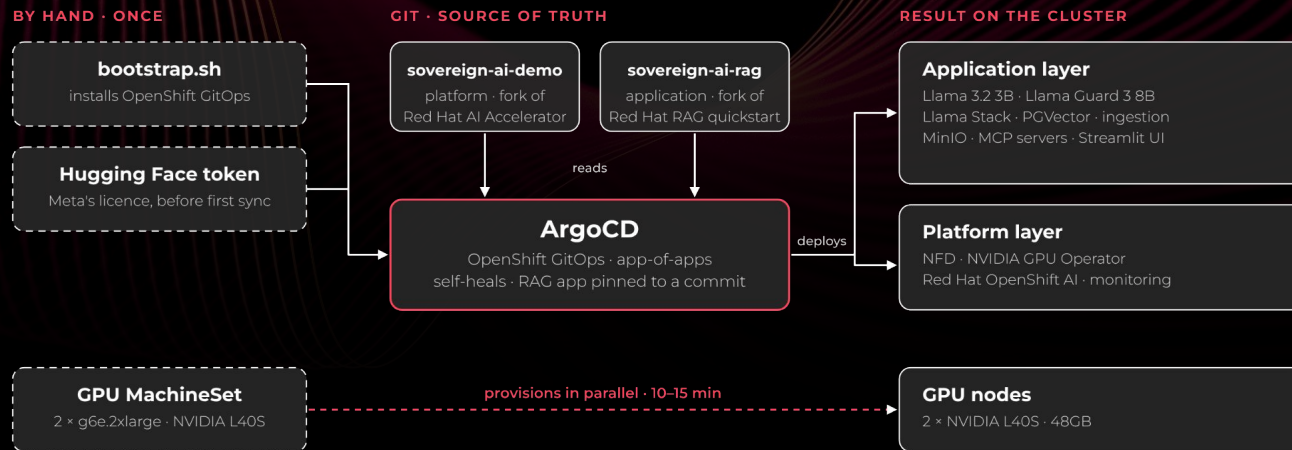




The Deployment Architecture

From empty cluster to running AI. Two repos. One GitOps bootstrap

Every query stays inside the cluster, from question to answer

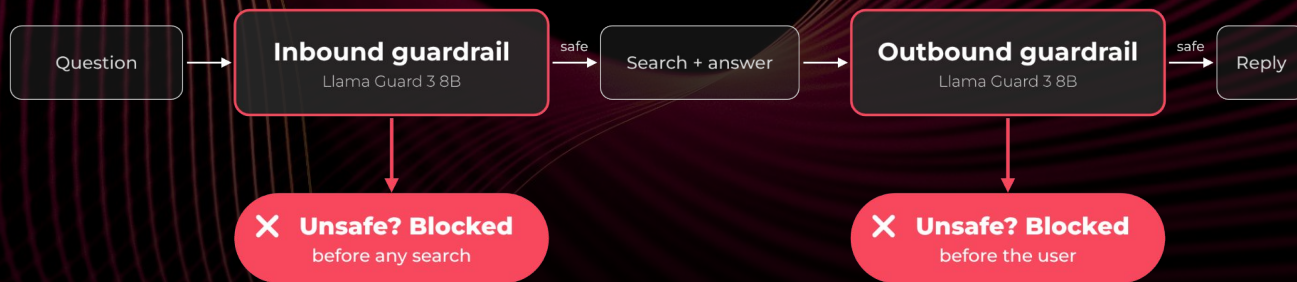




The AI Guardrail

Where governance meets runtime

Every message screened locally. Sovereign safety built into the Application layer





Serving: Il Primo - RAG





Lesson Learned

Two open source Red Hat repositories. One integration challenge.

Built on upstream. Wired together by Devoteam.



WHAT BROKE → HOW WE FIXED IT

- 1 The guardrail looked on, but wasn't**
→ Fixed in our fork, proven in pod logs
- 2 The default AI install didn't fit the cluster**
→ Minimal install: KServe, Llama Stack, GPU profiles
- 3 Healthy app, yellow dashboard**
→ Fixed Route drift and pipeline health check
- 4 The RAG app lived outside GitOps**
→ One GitOps tenant entry



The Agentic Researcher

Repeatable, evidence-based research

Which EU rules apply when FinanceCo grants a loan with AI?

Ask: "We want AI to decide who gets a loan. Which laws apply"

Sources: EU law from EUR-Lex, the web and internal documents

Outcome: Report · names the act and article, ready for expert review

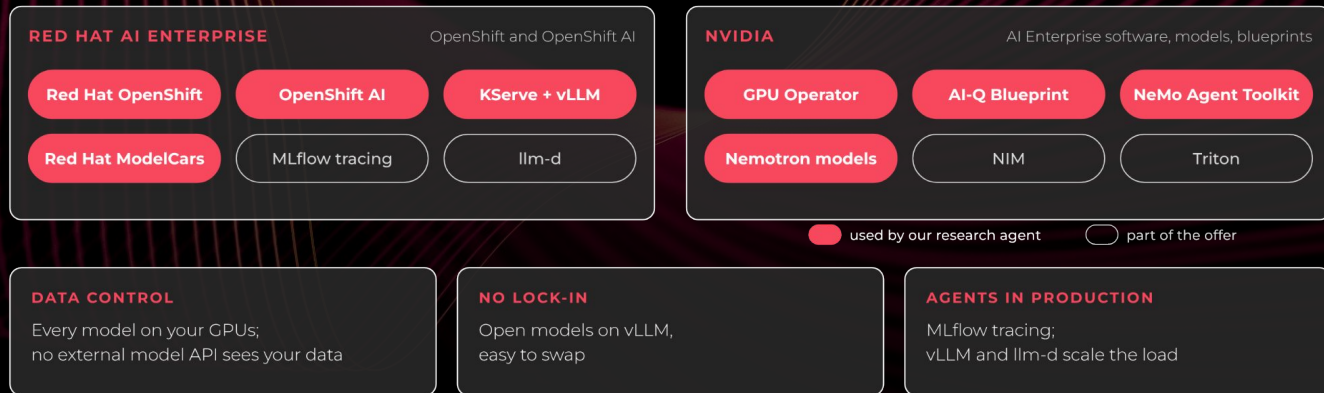
Constraint: models, documents and data stays in your cluster



Red Hat AI Factory with NVIDIA

NVIDIA brings the AI software. Red Hat brings the enterprise platform

Our research agent uses both halves of the offer

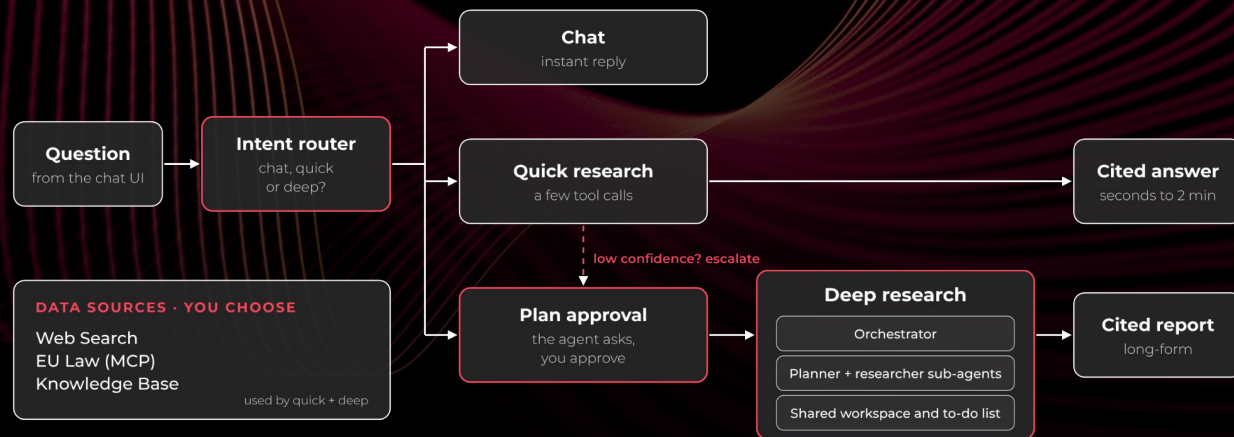




The Research Architecture

NVIDIA AI-Q · NeMo Agent Toolkit · LangChain Deep Agents

The agent decides how deep to go. You approve the plan

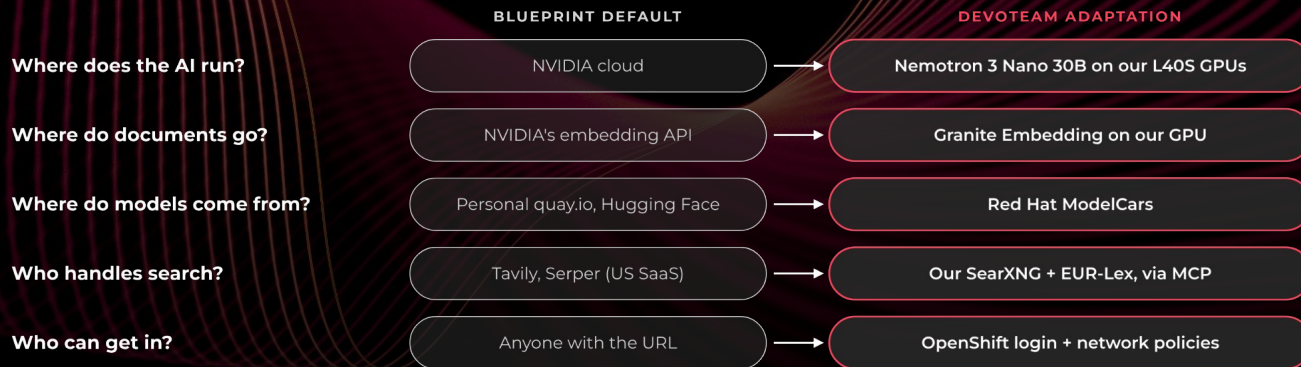




From Blueprint to Your Institution

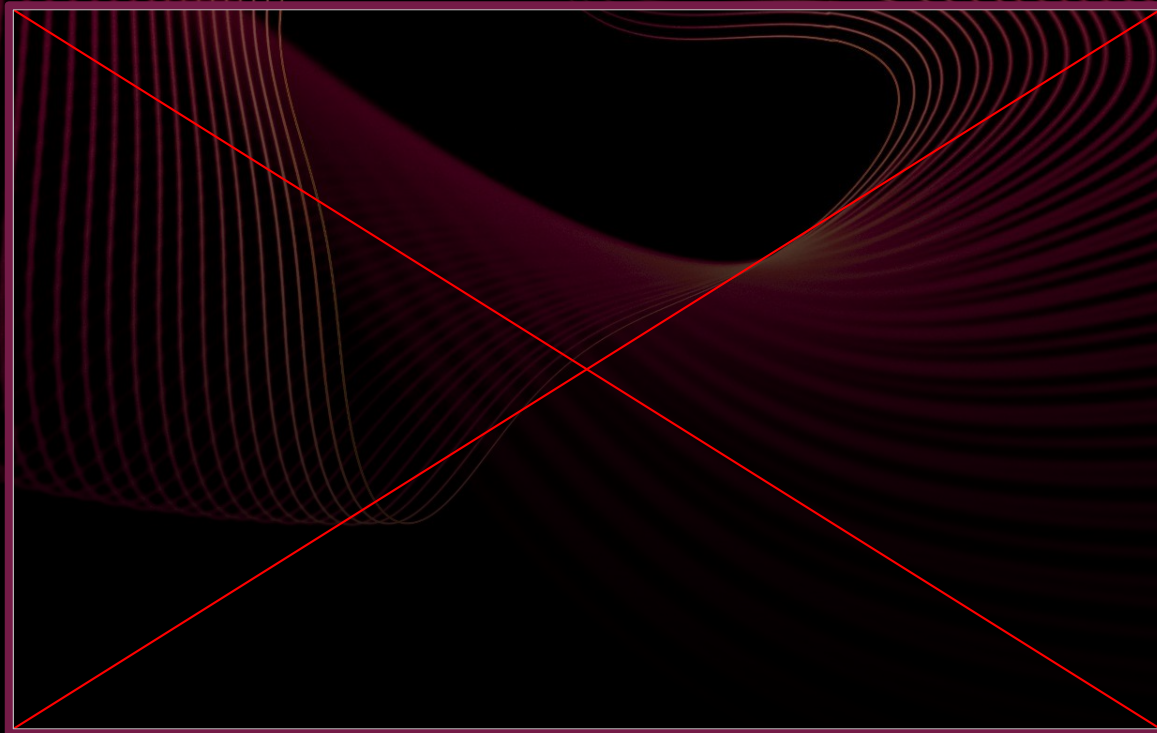
Devoteam adapts, integrates and runs AI for regulated customers

A cloud-first agent, rebuilt for our own infrastructure.





Serving: Il Secondo - Research





Lesson Learned

What worked, what didn't, what's next

The road to production is clear. Let's take it together



LESSON + IMPROVEMENT

- 1 Long research needs a long context window**
→ Bigger-memory GPUs
- 2 A 30B model gets legal details wrong**
→ Larger model; expert review stays
- 3 MCP made new sources easy to plug in**
→ Better EU Law citations: act name and link
- 4 Each blueprint solves one problem**
→ Add the RAG guardrail to the research agent
- 5 MCP servers are community images**
→ UBI-based images, deployed with GitOps
- 6 Models and tools change fast**
→ Automated evaluations for every change

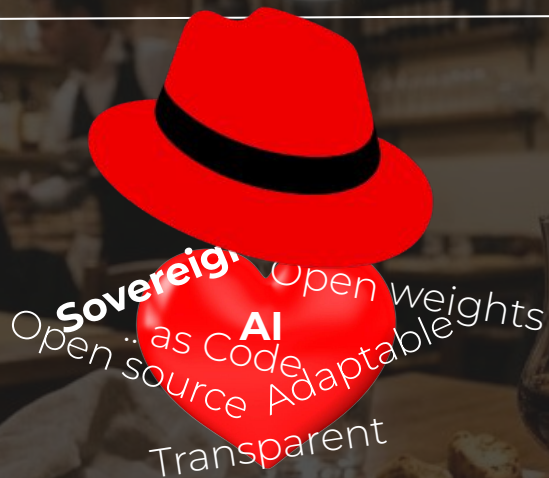


Il Formaggio e/o Il Dolce

Thank you for watching
Avatar Time



Il Caffè con Il Digestivo



Sovereign AI is no longer a concept. It has been built.

Join Us in the Kitchen

Let's cook something up together.



**Visit the
Devoteam Booth**



Fork the Repositories



**Solve Your
Challenges**

Buon Appetito!



devoteam