Sovereign AI: Building Independent AI Infrastructure with Open Source

CH-Open Open Source AI Conference
20 May 2026, Bern

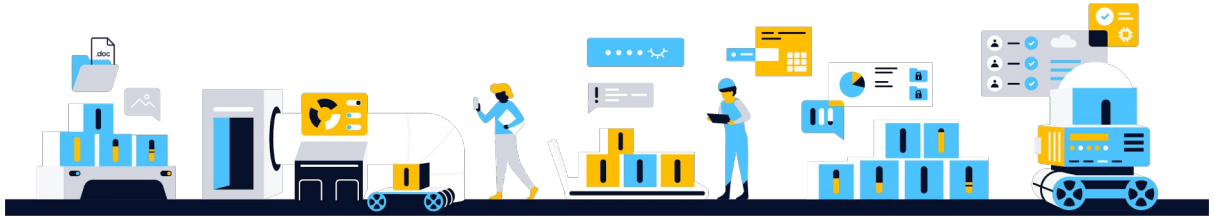
Aarno Aukia, co-founder
VSHN - The DevOps Company

Every Era Grew with “Open”

- Enterprise era: corp IT, then private cloud - Linux won because it was open
- Cloud era: public and hybrid cloud expanded with open source (Kubernetes, containers)
- AI era will be no different - but only if we build the infrastructure right from the start

AI Sovereignty Is Not Cloud Sovereignty

- Model provenance: who trained it, on what data, with what values baked in?
- Inference infra: who controls the serving layer and GPU access?
- Fine-tuning and RAG: your proprietary data enters the pipeline - where does it go?
- Lock-in at AI scale: token costs at production scale are enormous without portability



The EU Just Gave Us a Scoring Framework

- April 2026: EUR 180M in sovereign cloud contracts awarded by the EU Commission
- Providers scored on 8 sovereignty dimensions: SEAL-0 to SEAL-4 assurance levels
- 3x SEAL-3, 1x SEAL-2: Thales & Google joint venture cost Proximus a full SEAL level

vshn.ch: EUR 180M for Sovereign Cloud - what the EU procurement means for Swiss organisations

8 Dimensions of Sovereign AI Infrastructure

- Strategic (15%): Who owns the AI vendor? Any foreign parent or investors?
- Legal (10%): Which law governs your contracts? US CLOUD Act exposure?
- Data & AI (10%): Where is data processed? Who holds encryption keys for models and vector DBs?
- Operational (15%): Can your team fine-tune and patch without a foreign vendor?
- Supply Chain (20%): What's in your model stack, hardware, and sub-suppliers?
- Technology (15%): Is the AI stack open source? Can you migrate models and pipelines?
- Security (10%): Where is your AI-specific SOC? Who responds to a model poisoning incident?
- Environmental (5%): GPU cluster PUE and energy source - renewable or fossil?

VSHN Self-Assessment: SEAL-3 Equivalent

Dimension	Weight	Score	Key Evidence
Supply Chain	**20%**	**SEAL-3**	Open source stack; auditable BoM; Swiss / EU DCs
Strategic	15%	**SEAL-3**	VSHN AG Swiss-owned; no foreign parent or US investors
Operational	15%	**SEAL-3**	Team deploys, tunes & patches without foreign vendor
Technology	15%	**SEAL-3**	Kubernetes-native OS stack; portable workloads
Legal	10%	**SEAL-3**	Swiss law; GDPR; zero US CLOUD Act exposure
Data & AI	10%	**SEAL-3**	Data stays in CH; VSHN holds all encryption keys
Security	10%	**SEAL-3**	Own SOC; incident response; model integrity monitoring
Environmental	5%	**SEAL-3**	Swiss renewable energy; PUE ≤ 1.3

Sovereign AI Is a Stack You Assemble

- Layer 1 - Compute you control: Swiss cloud, on-prem GPUs, OpenShift / Kubernetes
- Layer 2 - Open models: Granite, Llama, Mistral - forkable, auditable, self-hostable
- Layer 3 - Sovereign inference: vLLM, llm-d, llmops.ch (VSHN)
- Layer 4 - Secure data and RAG: OpenBao, pgvector
- Layer 5 - Auditability: SBOMs, cryptographic software supply chain

Sovereignty by Model Type

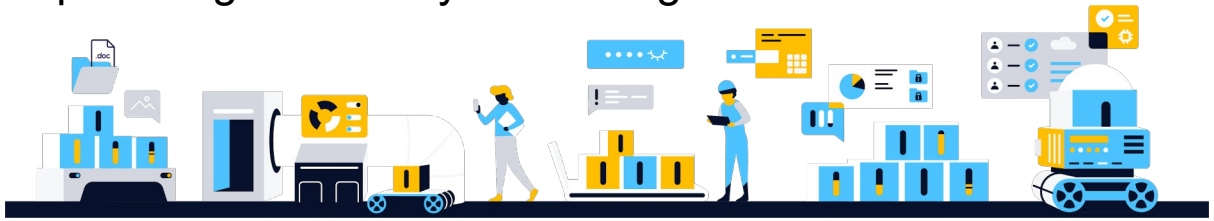
Dimension	Proprietary GPT-4o · Claude · Gemini	Open-Weight Llama 3 · Mistral · Phi-4	Open Source Granite · OLMo · Falcon
Strategic	✗ Foreign parent	⚠ License terms	✓ Community governed
Legal	✗ CLOUD Act exposed	⚠ Origin risk	✓ Open governance
Data & AI	✗ Vendor holds keys	⚠ Training data	✓ Full control
Operational	✗ API-only access	✓ Fine-tune locally	✓ Fork and patch
Supply Chain	✗ Opaque stack	⚠ Training opaque	✓ Full audit trail
Technology	✗ API lock-in	✓ Portable weights	✓ Fork + migrate
Security	✗ Shared vendor SOC	✓ Own your SOC	✓ Own your SOC
Environmental	✗ Opaque usage	⚠ Training	✓ Full transparency

Layer 2: Open Models - Own Your Weights

- Can it run air-gapped? Open weights: yes. Closed API: no - your AI stops if the vendor has an outage, changes terms, or export controls apply.
- Can you fine-tune it and keep the result? Closed APIs return a model ID, not weights - you rent access to your own tuned model. Open weights: the fine-tuned model is yours.
- Can you switch infrastructure without rewriting? Open models with standard OpenAI-compatible APIs let you move between providers. Proprietary SDKs lock you in.
- Do you know what it was trained on? Matters for IP risk, bias, and compliance. Open source: training data published. Open weight: weights only. Closed: neither.

Layer 3: Sovereign Inference

- vLLM: open source, high-performance LLM serving - runs on your own GPU infrastructure
- llm-d (Red Hat): distributed LLM inference for Kubernetes at scale - no cloud dependency
- llmops.ch (VSHN): managed LLM inference on the customer's own cluster - prompts and data never leave your control
- Hosted open inference (Infomaniak, Phoenix, Exoscale): token-based, no GPU needed - open weights mean you can migrate to self-hosted when ready



Layers 4 & 5: Secure Data and Auditability

- OpenBao: open-source secret and encryption key management for AI pipelines - your keys, your data
- PostgreSQL + pgvector: sovereign vector database for RAG - no proprietary vector DB lock-in
- Crossplane + Project Syn + K8up: infrastructure-as-code, GitOps audit trail, backup - full supply chain transparency
- Cryptographically signed software (sigstore, SBOMs): verify provenance of every component in your AI stack

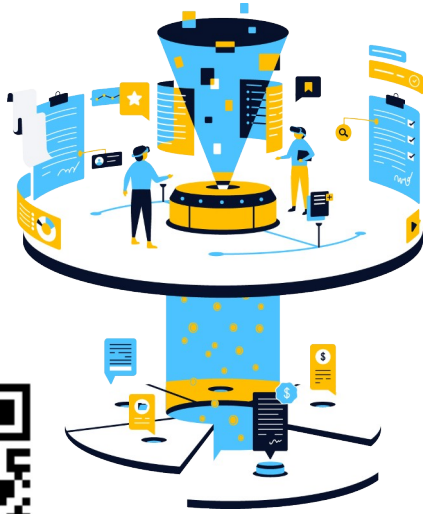
Infrastructure as a Bridge, Not a Bunker

- Sovereign infrastructure is not defensive - it grants you agility, compliance readiness, and AI freedom
- Open source transfers ownership from vendors to you - change providers, adopt new models, without rebuilding

Organizations on proprietary platforms ask permission.

Sovereign organizations ship.

Thank you and feedback very welcome!



Aarno Aukia

@aarnoaukia a@vs.hn

https://www.linkedin.com/in/aukia/