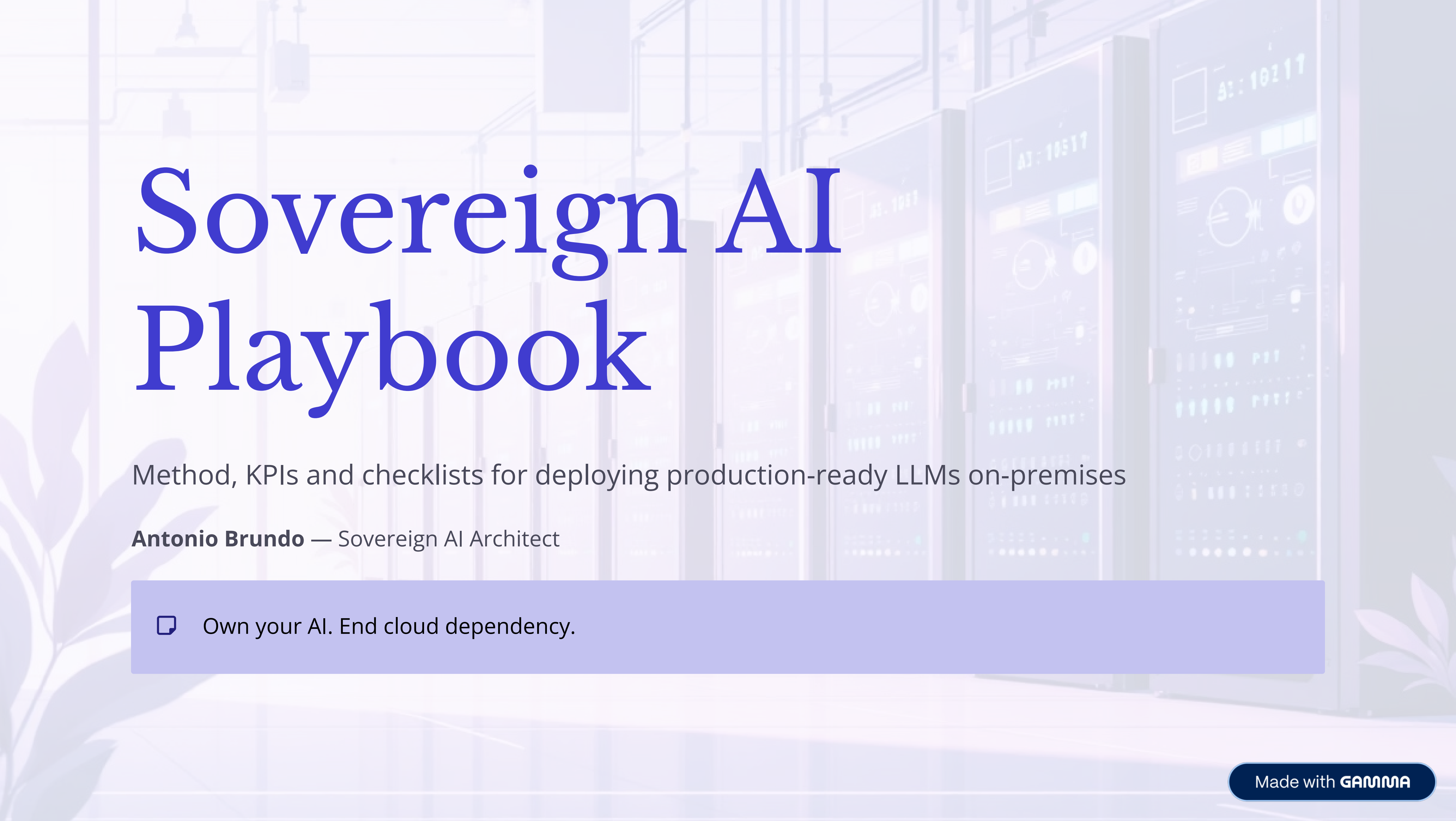


Sovereign AI Playbook

Method, KPIs and checklists for deploying production-ready LLMs on-premises

Antonio Brundo — Sovereign AI Architect

Own your AI. End cloud dependency.

Executive Summary

Large Language Models deployed on-premises offer a compelling value proposition when data sensitivity and predictable workloads align with sovereignty requirements. This approach delivers controlled latency, end-to-end auditability, and predictable Total Cost of Ownership whilst maintaining complete data governance.

Our proven five-step framework encompasses Data Governance, Model Strategy, Infrastructure, Security, and Operations & Governance. Each step includes clear Service Level Objectives and continuous measurement protocols to ensure optimal performance and compliance.

300ms

P95 Latency Target

Internal chat applications

25-60%

Cost Reduction

Per 1k tokens vs cloud baseline

0

Data Egress

On sensitive domains

This playbook provides decision makers with the practical tools, key performance indicators, and operational checklists needed to successfully deploy sovereign AI solutions that deliver measurable business value whilst maintaining complete control over sensitive data assets.

Sovereignty Principles

Sovereign AI implementation rests upon five fundamental principles that ensure complete control, transparency, and operational excellence. These principles guide every architectural decision and operational procedure.

Data Sovereignty

Zero uncontrolled egress with encryption by default and comprehensive audit trails. All data remains within defined boundaries with granular access controls.

Measurable Performance

Repeatable benchmarks tracking P50/P95 latency and throughput metrics. Performance transparency enables informed optimisation decisions.

Complete Verifiability

Full prompt-to-response traceability with automated evaluation pipelines. Every interaction is logged and measurable for quality assurance.

Unit Economics

Transparent cost per 1k tokens and per feature with built-in budget guardrails. Financial predictability through detailed cost modelling.

Strategic Portability

Reversible architectural choices with interchangeable components. Avoid vendor lock-in through standards-based implementation.

Sovereign AI Framework™: Step 1-2

1

Data Governance

Deliverables: Data mapping, access policies, ground truth datasets, and evaluation frameworks.

Critical Questions: Who accesses what data? Are personal data elements identified? What retention rules apply?

Anti-pattern: "Dumping everything into vector DB without classification."

KPIs: Test dataset coverage percentage, access incident count, compliance audit scores.

2

Model Strategy

Deliverables: Model family selection (Llama/Qwen), RAG/Fine-tuning plan, trade-off analysis table.

Critical Questions: Required context window? Tone and style requirements? Domain drift considerations?

Anti-pattern: "Defaulting to largest available model."

KPIs: Quality metrics (Exact/Partial match), hallucination percentage, cost per 1k tokens.

The first two steps establish the foundation for sovereign AI deployment. Data Governance ensures compliance and security whilst Model Strategy optimises for performance and cost-effectiveness. These steps must be completed thoroughly before infrastructure deployment begins.

Sovereign AI Framework™: Step 3-5

01

Infrastructure

Deliverables: Network topology (ingress→serving→vector), SLO definitions, resource quotas.

Key Decisions: Burst vs steady-state capacity, caching strategies, quantisation approaches.

KPIs: P95 latency, transactions per second, GPU/CPU utilisation, error rates.

02

Security

Deliverables: KMS/HSM integration, mTLS configuration, prompt injection policies, SIEM feeds.

Critical Areas: Attack surface analysis, CISO audit requirements, compliance frameworks.

KPIs: Mean Time To Recovery, key rotation percentage, log coverage metrics.

03

Operations & Governance

Deliverables: Evaluation pipeline, canary deployment, versioning strategy, operational runbooks.

Decision Criteria: Promotion and rollback triggers, change management processes.

KPIs: Lead time to change, change failure rate, drift benchmark scores.

Steps 3-5 operationalise the sovereign AI platform with robust infrastructure, comprehensive security, and mature operational practices. The anti-pattern to avoid is "set and forget" mentality—continuous monitoring and improvement are essential for success.

Key Architectural Decisions

Three critical decision trees guide architectural choices, each with specific triggers and trade-offs that impact performance, cost, and operational complexity.

RAG vs Fine-Tuning

RAG: Ideal for changing domains and compliance requirements

Fine-Tuning: Optimal for consistent style and extreme latency needs

Hybrid: Best when both flexibility and performance matter

Latency vs Quality

Sub-300ms: Compact models, distillation, KV caching

High Quality: Request batching with intelligent caching

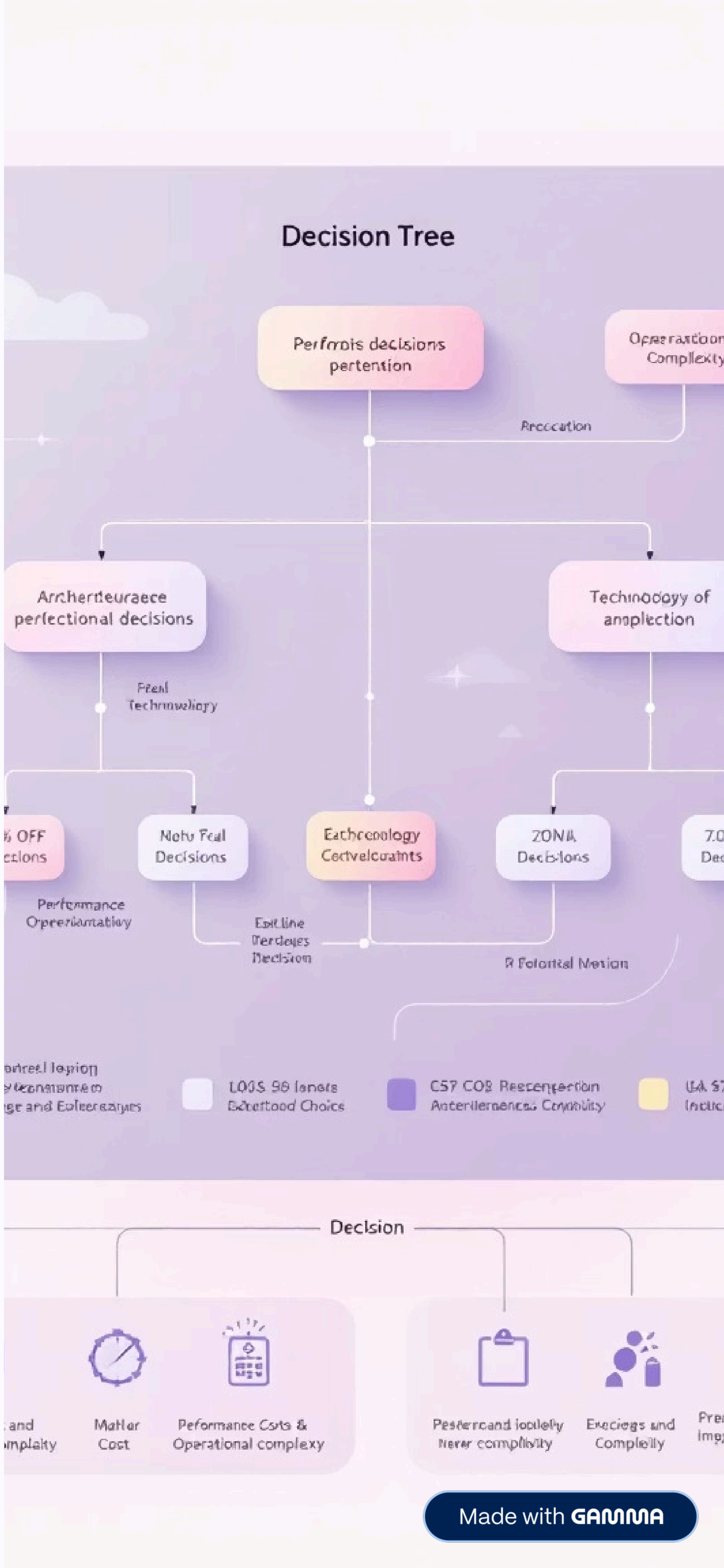
Balanced: Dynamic model routing based on query complexity

Vector Database Selection

Embedded: Small datasets with simple access patterns

Milvus/PGVector: Scale requirements with ACL complexity

Hybrid: Multi-tier storage for performance optimisation


Decision Tree
Perfrois decisions pertention
Operatioon Complexity
Rreccation
Archerteuraece perfectional decisions
Techinodogy of anplection
Feeal Techmmalory
OFF cions
Nofu Feal Decisions
Eathcenology Cectvelcaants
ZONA Decislons
Performance Opperiantatlory
Exit line Terdeyes Decision
R Folortal Metion
LO$S 98 Ianets Bateftood Choice
C57 CO8 Recenpection Antenlernences Crmmility
Decision
Matter Cost
Pefonmance Csts & Operational complexy
Pesterrcand loolelly Nerer complivity
Eveciegs and Complelly

KPIs & Maturity Assessment

Comprehensive measurement framework tracking operational excellence across five maturity levels, from initial exploration to full sovereignty.

KPI Category	Metric	Target Range	Frequency
Performance	P50/P95 Latency	< 150ms / < 300ms	Real-time
Cost	Cost per 1k tokens	25-60% below cloud	Daily
Quality	Hallucination rate	< 5%	Continuous
Reliability	Error rate	< 0.1%	Real-time
Adoption	User engagement	> 70% monthly active	Weekly

1 **L0 Exploration**

Proof of concept with basic functionality

2 **L1 Pilot**

Limited production deployment

3 **L2 Production**

Full operational deployment

4 **L3 Industrialised**

Automated operations at scale

5 **L4 Sovereign**

Complete independence and control

Operational Checklists

Comprehensive pre-pilot and go-live checklists ensuring nothing is overlooked during critical deployment phases. Each item includes space for notes and sign-off.

Pre-Pilot Checklist

- Data classification completed
- Access controls defined
- Model selection validated
- Infrastructure provisioned
- Security policies implemented
- Evaluation metrics established
- Monitoring dashboards configured
- Backup procedures tested
- Team training completed
- Stakeholder sign-off obtained

Go-Live Checklist

- Performance benchmarks met
- Security audit passed
- Load testing completed
- Disaster recovery verified
- User documentation ready
- Support procedures active
- Scaling policies defined
- Compliance validation done
- Rollback plan tested
- Go-live approval confirmed

These checklists serve as quality gates ensuring systematic progression through deployment phases. Regular review and updates based on lessons learned maintain their effectiveness over time.

ROI & Unit Economics

Comprehensive financial framework for evaluating sovereign AI investments, incorporating productivity gains, quality improvements, and risk mitigation benefits against capital and operational expenditure.

> ## ROI Formula
>
> **ROI = (ΔProductivity + ΔQuality + Risk Avoided) - (CapEx/Amortisation + OpEx)**

Productivity Gains

- Reduced manual processing time
- Accelerated decision-making cycles
- Automated routine tasks
- Enhanced knowledge discovery

Quality Improvements

- Reduced human error rates
- Consistent output quality
- Enhanced compliance accuracy
- Better customer satisfaction

Risk Mitigation

- Data breach prevention
- Regulatory compliance assurance
- Vendor dependency reduction
- Operational resilience

Example Calculation: A mid-size organisation might see £500K annual productivity gains, £200K quality improvements, and £300K risk mitigation against £400K total costs, yielding 150% ROI. Sensitivity analysis should examine key variables including adoption rates, performance targets, and cost escalations.

Contact & Next Steps

Direct Contact

antonio.brundo@sovereign-ai.com

PGP key available upon request

Consultation Booking

Scan QR code for calendar access

Initial assessment sessions available

Privacy Notice

All communications encrypted

NDA available for sensitive discussions

Frequently Asked Questions

> **Q: What's the typical implementation timeline?**
>
> Most organisations achieve L2 production readiness within 3-6 months, depending on data complexity and security requirements.

> **Q: How do costs compare to cloud alternatives?**
>
> On-premises deployment typically reduces per-token costs by 25-60% whilst eliminating data egress fees and providing predictable scaling costs.

Ready to begin your sovereign AI journey? Contact us for a confidential assessment of your requirements and a customised implementation roadmap tailored to your organisation's specific needs and constraints.