Sovereign AI On-Premise: Measured Results on Latency, Quality and TCO

This compendium presents three comprehensive case studies demonstrating the tangible benefits of implementing sovereign AI solutions on-premise within enterprise environments. Each case study provides detailed analysis of performance metrics, total cost of ownership (TCO) reductions, and quality improvements achieved through strategic deployment of large language models in controlled, secure environments.

The following analyses showcase real-world implementations across finance, telecommunications, and pharmaceutical sectors, offering decision-makers concrete evidence of sovereign AI's capacity to deliver measurable business value whilst maintaining complete data sovereignty and regulatory compliance.

Understanding the Methodology

Service Level Objectives (SLOs)

All performance metrics are measured against predefined SLOs including p95 latency thresholds, availability targets of 99.9%, and quality benchmarks established through controlled testing environments.

Key Performance Indicators

KPIs encompass operational metrics (latency, throughput), financial metrics (TCO, cost per token), quality metrics (hallucination rates, factuality scores), and adoption metrics (user engagement, deflection rates).

Data Anonymisation

All case studies have been carefully anonymised to protect client confidentiality whilst preserving the integrity of performance data. Sensitive business information remains on-premise and is never exposed to external systems.

These case studies represent controlled implementations where baseline measurements were established prior to deployment, enabling accurate delta calculations. Each implementation followed rigorous testing protocols to ensure data reliability and reproducibility within similar enterprise contexts.

The metrics presented reflect internal measurements on controlled datasets. Replicability is subject to specific organisational context, existing infrastructure capabilities, and implementation methodology. All financial calculations include infrastructure, operational, and licensing costs to provide comprehensive TCO analysis.

Finance Sector: Banking Knowledge Assistant

220ms
p95 Latency
Response time improvement

37%
TCO Reduction
Total cost savings achieved

1.8%
Hallucination Rate
Quality improvement metric

1.8K
Monthly Users
Adoption rate achieved

Executive Summary

The implementation delivered a sophisticated knowledge assistant leveraging Retrieval-Augmented Generation (RAG) with Access Control Lists (ACL), comprehensive audit trail capabilities, and dramatic cost reduction from €0.024 to €0.006 per 1,000 tokens. The on-premise deployment ensures complete data sovereignty whilst delivering enterprise-grade performance and reliability.

KPI	Before	After	Delta
Latency p95	1,200 ms	220 ms	-980 ms
Cost per 1k tokens	€0.024	€0.006	-75%
Hallucination rate	10.0%	1.8%	-8.2 pt
Knowledge coverage	62%	91%	+29 pt
User adoption/month	0	1,800	+1,800

Finance Implementation: Technical Architecture

Weeks 1-2: Assessment

1

Infrastructure evaluation, requirements gathering, and baseline establishment. Technical feasibility analysis and security framework design.

2

Weeks 3-5: Pilot Phase

Limited deployment with selected user groups. Performance tuning, model optimisation, and integration testing with existing banking systems.

Week 6: Production

3

Full production deployment with comprehensive monitoring, alerting, and backup procedures. Complete documentation and knowledge transfer.

Key Technical Decisions

The architecture prioritised RAG over fine-tuning to maintain agility and reduce training overhead. Implementation of vLLM with key-value caching significantly improved inference efficiency. Milvus vector database with ACL provides granular access control aligned with banking compliance requirements.

Technology Stack: 13B parameter model with INT4 quantisation, e5-small embeddings, 4×H200 GPU configuration, integrated with Key Management Service (KMS), Hardware Security Module (HSM), and Security Information and Event Management (SIEM) systems.

> "We can now measure costs and quality on a daily basis, providing unprecedented visibility into our AI operations whilst maintaining complete regulatory compliance." — Head of Operations

Telecommunications: NOC Ticket Triage

180ms
p95 Latency
Response optimisation

41%
L1 Deflection
Automated resolution rate

32%
MTTR Reduction
Mean time to resolution

24%
TCO Reduction
Operational cost savings

Performance Transformation

The telecommunications implementation focused on intelligent ticket triage within the Network Operations Centre (NOC), achieving remarkable improvements in operational efficiency. The system processes incoming support tickets, categorises them by severity and type, and provides immediate resolution guidance for Level 1 support staff.

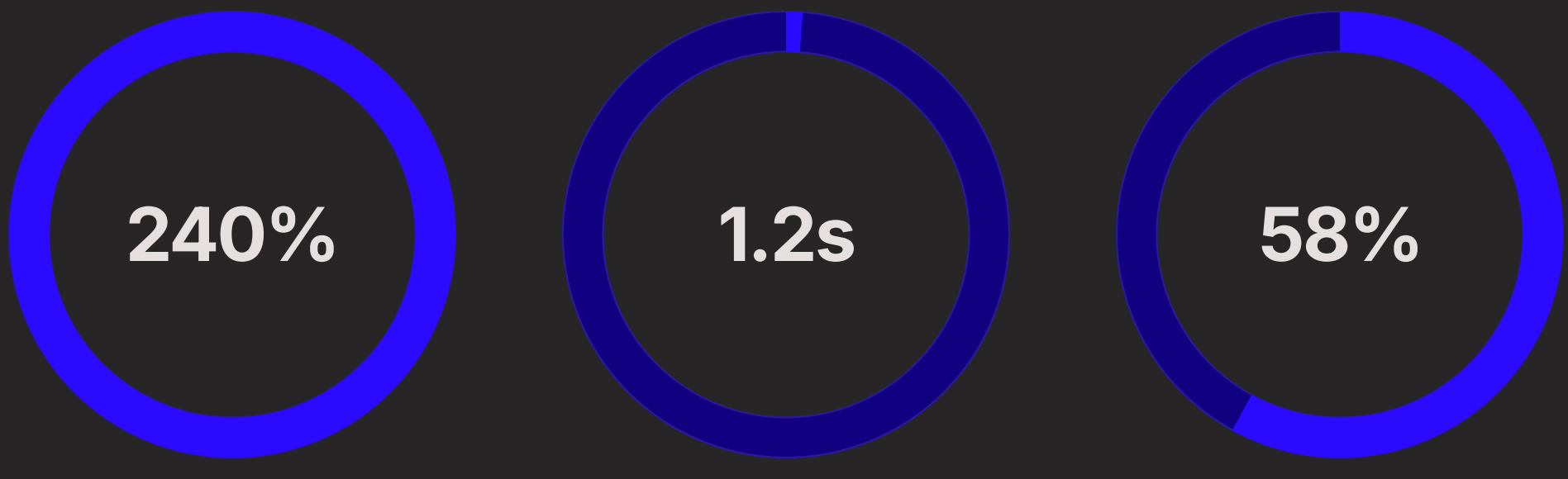
KPI	Before	After	Delta
Latency p95	420 ms	180 ms	-240 ms
L1 Deflection	0%	41%	+41 pt
MTTR	Baseline	-32%	-32%
Cost per ticket	Baseline	-€0.78	-€0.78

Architecture Decisions

The solution employs an 8-13B parameter model optimised for telecommunications domain knowledge. RAG implementation prioritises recent documentation to ensure currency of information. Speculative decoding provides significant performance improvements for typical NOC query patterns.

Technical Stack: vLLM inference engine, migration from FAISS to Milvus for enhanced scalability, proprietary in-house embedding models trained on telecommunications documentation, integrated with existing ITSM platforms.

Pharmaceutical: R&D Documentation Summarisation



Throughput Increase
Document processing capacity improvement

p95 Latency
Response time for complex summaries

Cost Reduction
Processing cost per 1k tokens

Research Documentation Excellence

The pharmaceutical implementation addresses the critical challenge of processing vast quantities of research documentation, clinical trial data, and regulatory submissions. The solution provides intelligent summarisation whilst maintaining factual accuracy essential for pharmaceutical research environments.

KPI	Before	After	Delta
p95 Latency	4.3 s	1.2 s	-3.1 s
Cost per 1k tokens	Baseline	-58%	-58%
Factuality Score	72.1	81.5	+9.4 pt

Technical Implementation

The architecture combines lightweight fine-tuning for pharmaceutical writing style with RAG for maintaining up-to-date information. Mixed CPU-GPU batch processing optimises resource utilisation for varying document sizes and complexity levels.

Technology Foundation: TensorRT-LLM for optimised inference, PGVector for scalable vector storage, bi-encoder architecture with 768-dimensional embeddings specifically tuned for scientific literature processing.

Key insight: BM25 combined with embedding search proves particularly robust for lengthy specialised texts common in pharmaceutical research, delivering superior relevance compared to embedding-only approaches.

Comparative Analysis Across Sectors

1

Finance Sector

- Ultra-low latency: 220ms p95
- Highest TCO reduction: 37%
- Exceptional quality: 1.8% hallucination
- Strong adoption: 1,800 users monthly

2

Telecommunications

- Fastest response: 180ms p95
- Solid TCO gains: 24% reduction
- Operational impact: 41% deflection
- MTTR improvement: 32% reduction

3

Pharmaceutical

- Complex processing: 1.2s p95
- Highest cost savings: 58%
- Quality improvement: +9.4 factuality
- Throughput gains: 240% increase

Sector	p95 Latency	TCO Reduction	Quality Metric	Key Success
Finance	220 ms	-37%	1.8% hallucination	1.8k users
Telco	180 ms	-24%	Quality maintained	41% deflection
Pharma	1.2 s	-58%	+9.4 pt factuality	240% throughput

Cross-Sector Insights

Each implementation demonstrates distinct optimisation patterns aligned with sector-specific requirements. Finance prioritises speed and accuracy for customer-facing applications, telecommunications focuses on operational efficiency and deflection rates, whilst pharmaceutical emphasises processing capacity and factual precision for research applications.

The consistent achievement of significant TCO reductions across all sectors validates the economic viability of sovereign AI implementations. Performance improvements compound over time as systems learn and optimise based on usage patterns specific to each organisation's operational context.

Reusable Implementation Template

Essential Framework Components

01

Requirements Assessment

Infrastructure evaluation, performance requirements definition, compliance framework mapping, and baseline metric establishment across operational, financial, and quality dimensions.

02

Architecture Design

Technology stack selection, model sizing decisions, integration planning with existing systems, security framework implementation, and monitoring strategy development.

03

Pilot Implementation

Controlled deployment with limited user base, performance tuning and optimisation, quality assurance testing, and feedback integration for refinement.

04

Production Deployment

Full-scale rollout with comprehensive monitoring, operational procedures documentation, staff training delivery, and continuous improvement processes.

Standard Deliverables Framework

Technical Documentation

- Reference architecture diagrams
- Request and ingestion flow documentation
- Operations and CI/CD procedures
- Security control matrices by zone
- Performance benchmarking results

Business Documentation

- Executive summary with key metrics
- Before/after performance comparison
- Implementation timeline and milestones
- Lessons learned and recommendations
- ROI analysis and cost projections

This template approach ensures consistency across implementations whilst allowing for sector-specific customisation. Each component has been validated through multiple deployments and incorporates best practices derived from successful enterprise implementations across diverse industry verticals.

Technical Architecture Reference

Core Infrastructure Components

Inference Layer

High-performance GPU clusters with vLLM or TensorRT-LLM optimisation, supporting model parallelism and batched inference for maximum throughput efficiency.

Vector Storage

Scalable vector databases (Milvus, PGVector) with ACL integration, supporting semantic search, hybrid retrieval, and real-time index updates.

Security Framework

Comprehensive security controls including KMS/HSM integration, SIEM monitoring, audit trails, and zone-based access controls ensuring regulatory compliance.

Orchestration

Container orchestration with Kubernetes, automated CI/CD pipelines, monitoring and alerting systems, and disaster recovery procedures for operational resilience.

Data Flow Architecture

The reference architecture supports both real-time inference and batch processing workflows. Request routing optimises between cached responses and live model inference based on query patterns and performance requirements. Ingestion pipelines handle document processing, embedding generation, and vector storage updates with configurable refresh cycles.

Security controls operate at multiple layers: network segmentation isolates AI workloads, application-level ACLs enforce fine-grained permissions, and audit systems track all interactions for compliance reporting. This layered approach ensures enterprise-grade security without compromising performance or usability.

Strategic Assessment Opportunity

Next Steps for Your Organisation

The case studies presented demonstrate the tangible value achievable through strategic sovereign AI implementation. Each organisation's journey begins with a comprehensive assessment to identify optimal use cases, quantify potential benefits, and design an implementation roadmap aligned with specific business objectives and technical constraints.

Strategic Assessment

Comprehensive evaluation of current infrastructure, identification of high-value use cases, and development of business case with projected ROI and implementation timeline.

Implementation Planning

Detailed technical architecture design, resource requirement specification, timeline development, and risk mitigation strategy formulation for successful deployment.

Deployment Excellence

Guided implementation with best practices application, performance optimisation, staff training delivery, and establishment of operational procedures for ongoing success.

Contact Information

To explore how sovereign AI can transform your organisation's operations whilst maintaining complete data sovereignty, we invite you to participate in a Strategic Assessment. This comprehensive evaluation will identify your organisation's optimal AI implementation strategy and quantify the potential business value achievable through on-premise deployment.

Antonio Brundo

Senior AI Strategist & Implementation Specialist

Specialising in enterprise sovereign AI solutions with proven track record across finance, telecommunications, and pharmaceutical sectors.

Assessment Request: Contact us to schedule your strategic evaluation and receive a customised implementation roadmap for your organisation.

Scan to access contact portal and schedule your Strategic Assessment consultation.