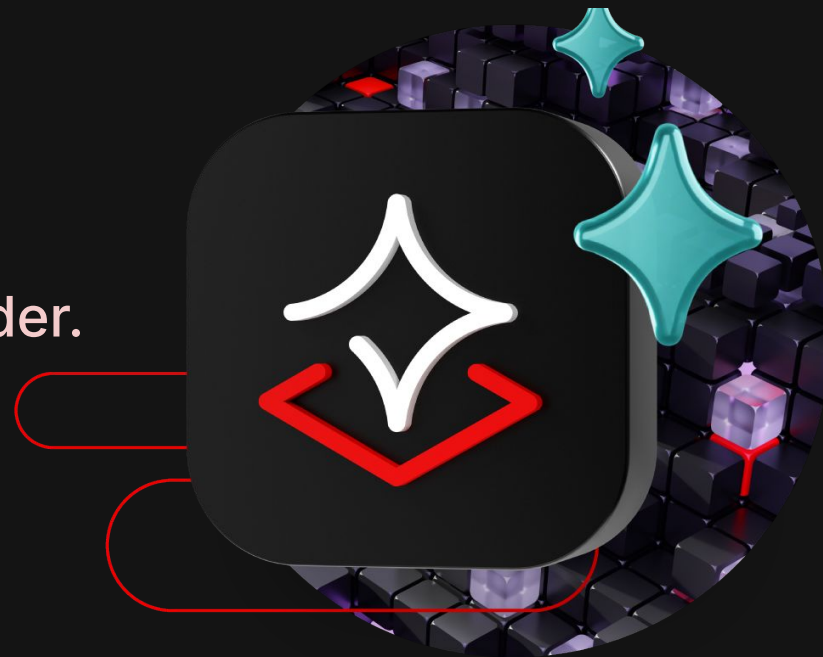


The AI Paradox

Prototyping is faster than ever, but shipping is harder.

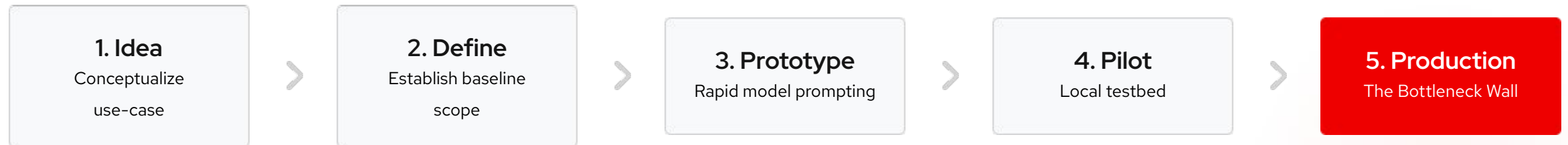
Brendan OHara
Chief Architect
brohara@redhat.com



THE SANDBOX BOTTLENECK

| THE ILLUSION OF EASE

Why teams start highly confident, tracing the "happy path" of isolated sandboxes and public APIs until they hit the enterprise production bottleneck.



KEY TAKEAWAY: IF AI IS SO EASY, WHY DO MOST TEAMS GET STUCK BETWEEN PILOT AND PRODUCTION?

STRATEGIC DILEMMA

| THE DAY-ONE GOVERNANCE GAP

Why they hit that bottleneck—the "Day-One Governance Gap" where teams sacrifice long-term control for short-term demo speed.



The Speed Bias (Day 1)

Teams optimize exclusively for immediate demo speed. They download arbitrary open models, ignore data boundaries, and bypass standard configuration parameters to show early results fast.



The Operational Legacy (Day 2)

When the project is ready to scale, Infosec, Compliance, and Operations step in. Without strong governance choices made early, the architecture must be completely scrapped and rewritten from scratch.

KEY TAKEAWAY: PRIORITIZING SHORT-TERM SPEED OVER LONG-TERM CONTROL IS A TRAP.

THE ENGINEERING EPIPHANY

| THE 20 / 80 REALITY SPLIT

20%

MODEL SELECTION & PROMPTING

Integrating pre-trained models and writing prompt logic is as little as 20% of the effort. The remaining 80% is enterprise application architecture: RAG pipelines, agent orchestration, access controls, and operational runtimes.

For company after company deploying, governing, scaling, and safeguarding the surrounding infrastructure layers consumes the bulk of operational lifecycle efforts.

KEY TAKEAWAY: STOP REINVENTING INFRASTRUCTURE - ENTERPRISE VALUE DEPENDS ON OPERATIONAL RUNTIMES.

A PRAGMATIC AI STRATEGY

| PUBLIC APIS & PRIVATE MODELS MUST CO-EXIST

Your strategy should complement public cloud APIs by providing the governance and runtime controls needed to safely run private models alongside cloud endpoints for maximum flexibility.

Frontier Public Cloud APIs

Examples: Gemini, OpenAI, Claude

- Unconstrained multi-modal tasks & complex reasoning
- Rapid ideation, exploratory research, and prototyping
- Broad general knowledge queries across public domains
- **Focus:** Instant access to cutting-edge general capabilities

Private Domain Models

Engineered with Red Hat OpenShift AI

- Air-gapped regulated workloads & sovereign compliance
- Sensitive customer PII, trade secrets, and financial data
- High-volume, repetitive enterprise domain tasks
- **Focus:** Maximum control, data privacy, & predictable unit cost

KEY TAKEAWAY: ENTERPRISE AI IS NOT AN EITHER/OR CHOICE. MATCH WORKLOADS TO THE OPTIMAL ARCHITECTURE.

PROBLEM #1: API COUPLING & SDK VOLATILITY

| ARCHITECTURAL ABSTRACTION

Large Language Model improvements mean the foundation layer is shifting daily, requiring architectural abstraction to prevent structural technical debt.

The Shifting Baseline

Rapid shifts in context limits, prompt structures, and proprietary API standards constantly destabilize live applications.

The Feature Coupling Penalty

Building application logic directly against provider-specific SDKs locks your product roadmap into a single ecosystem.

The Overhaul Penalty

Adopting standard model gateways allows you to route queries to public APIs or swap in private open models without rewriting orchestration code.

KEY TAKEAWAY: NOBODY WANTS TO REWRITE THEIR APPLICATION EVERY SIX MONTHS JUST TO STAY CURRENT.

SOLUTION 1: HARDENED AGENT HARNESS & UNIFIED GATEWAYS

| STABILITY, FLEXIBILITY, AND SECURE INNOVATION



THE SOLUTION

Decouple application logic using standard model gateways (MaaS) while wrapping open-source agent frameworks inside a secure, enterprise-hardened harness.



KEY ENABLER

Red Hat OpenShift AI provides an OpenAI-compatible API Gateway alongside an **Enterprise Agent Harness** to isolate agent runtimes in zero-trust sandboxes with automated AI Evals



BUSINESS IMPACT

Safely adopt cutting-edge open-source AI innovation and swap base models in the background without modifying core application code or violating enterprise risk standards.

KEY TAKEAWAY: DECOUPLE APPLICATION LOGIC AND HARDEN OPEN INNOVATION TO MAINTAIN ABSOLUTE CONTROL

PROBLEM #2: BRITTLE DATA PLUMBING

| THE BYO INFRASTRUCTURE TRAP

80%

OF BUILDER TIME UNDERUTILIZED

Developers are forced to spend their critical operational cycles building fragile plumbing instead of driving true core business value.

Developers waste cycles writing ad-hoc document parsers, stitching together unsynced vector stores, cleaning unstructured data, while lacking standardized mechanisms to generate high-quality synthetic datasets for model alignment.

KEY TAKEAWAY: WE ARE FORCING HIGHEST-VALUE BUILDERS TO ACT AS BASIC PLUMBING ASSISTANTS.

SOLUTION 2: AUTOMATED DATA PIPELINES & SYNTHETIC DATAGATEWAYS

| STABILITY AND FLEXIBILITY



THE SOLUTION

Automate foundational data pipelines end-to-end, standardizing data prep and synthetic model alignment.



KEY ENABLER

SDG Hub (Synthetic Data Generation Hub) and **OpenShift AI Pipelines** automate high-quality synthetic data generation, allowing domain experts to easily align models with business logic.



BUSINESS IMPACT

Empower domain experts to fine-tune AI models using raw business knowledge, removing the need for a PhD in Data Science.

KEY TAKEAWAY: EMPOWERS DOMAIN EXPERTS TO SAFELY INJECT BUSINESS KNOWLEDGE.

PROBLEM #3: FRAGMENTED GPU SILOS

| COMPUTE RESOURCE STARVATION

Hardware hoarding and siloed compute leave active development teams starved for GPU resources.



Isolated Data Science Hubs

Individual data science teams may hoard dedicated physical GPU pools. Audits frequently reveal isolated hardware sitting idle up to 80% of the time because infrastructure is heavily siloed and unshared.



Developer Resource Starvation

Downstream software engineering and application teams therefore remain completely locked out of compute access, bottlenecking production code validation and integration.

KEY TAKEAWAY: DISTRIBUTED BUT SILOED CLUSTERS MAY SIT IDLE WHILE DEVELOPERS STARVE.

SOLUTION 3: UNLOCKING FRAGMENTED COMPUTE

| DYNAMIC HARDWARE QUOTAS



THE SOLUTION

Establish dynamic multi-tenant hardware quotas to safely coordinate and balance complex GPU tasks across active business teams.



KEY ENABLER

Kueue Workload Scheduler provides a native Kubernetes engine for active resource pooling and automated GPU sharing.



BUSINESS IMPACT

Maximize enterprise GPU physical ROI, driving up hardware utility rates and ensuring zero waste on costly processing infrastructure.

KEY TAKEAWAY: USE POOLING TO MAXIMIZE ACTIVE UTILIZATION OF PHYSICAL GPUS.

PROBLEM #4: SHADOW AI FLEETS

| THE COMPLIANCE RISK

Ungoverned agents and hardcoded credentials create severe compliance and data exposure risks.



Broken Agent Identity & "YOLO Mode" Access

Traditional IAM tools break down in agentic environments. Agents either run in unrestricted "YOLO mode" (inheriting full user entitlements) or inadvertently discover API credentials across firewalls.



Ungoverned Autonomy

Security teams face catastrophic data leakage and compliance exposures because unknown agent code execution behaviors remain completely invisible to IT oversight.

KEY TAKEAWAY: SECURITY TEAMS SAY "NO" TO PRODUCTION BECAUSE THEY HAVE ZERO VISIBILITY.

SOLUTION 4: GOVERNING SHADOW AI FLEETS

CREDENTIAL ISOLATION & DETERMINISTIC OVERSIGHT



THE SOLUTION

Execute AI workflows within secure, credential-isolated runtime nodes while enforcing mandatory, audited **Human-in-the-Loop (HITL)** approval checkpoints for sensitive operations.



KEY ENABLER

MaaS API Gateway & Centralized Secrets Management enforce zero-trust credential isolation. **Deterministic Approval Workflows** for human oversight. **MCP** & **OGX** provide standard interfaces.



BUSINESS IMPACT

Provide information risk officers and CISOs with a unified, tamper-resistant control system that transforms unmonitored agent risk into an auditable, enterprise-compliant approval flow.

KEY TAKEAWAY: ISOLATION, CREDENTIAL DELEGATION, AND HUMAN OVERSIGHT NEUTRALIZE COMPLIANCE RISKS

PROBLEM #5: OPAQUE AI OPERATIONS

| AI BREAKS DIFFERENTLY

APM tools say the server is fine, because legacy monitoring tools have a critical blind spot.

Traditional Software

- CPU Performance & Allocation
- Disk Space & Memory IOPS
- Network Latency Curves

Volatile GenAI Systems

- Hallucinations & Alignment Drift
- Token Queue Depths & KV Cache Pressures
- Time-to-First-Token (TTFT) Degradations

KEY TAKEAWAY: ENTERPRISE AI BREAK COMPLETELY DIFFERENTLY THAN TRADITIONAL ENTERPRISE APPLICATIONS

SOLUTION 5: OPERATIONAL TRANSPARENCY

| SPECIALIZED AI MONITORING



THE SOLUTION

Deploy specialized GenAI observability and real-time evaluation telemetry designed strictly for live, production-grade LLMs and multi-agent workflows.



KEY ENABLER

Centralized AI Monitoring & Observability deliver real-time telemetry tracking safety guardrails and output alignment alongside token-level metrics like TTFT and queue depth.



BUSINESS IMPACT

Empower SRE and Platform teams to immediately catch output regressions, enforce compliance guardrails, and auto-scale infrastructure based on actual token queue pressure.

KEY TAKEAWAY: GENAI TELEMETRY TURNS PROTOTYPE SPEED INTO ENTERPRISE PRODUCTION.

PROBLEM #6: PRODUCTION WORKLOAD ECONOMICS

| SCALING AT ENTERPRISE VOLUME

Predictable pilot workloads require workload-optimized infrastructure as production adoption scales.

Monolithic Container Bloat

Rapid shifts in context limits, prompt structures, and proprietary API standards constantly destabilize live applications.

Workload Scale Economics

Building application logic directly against provider-specific SDKs locks your product roadmap into a single ecosystem.

The Token Cost Wall

High-volume tasks cause token spends to explode, burning through annual budgets in months and forcing emergency budget reviews.

KEY TAKEAWAY: HIGH-VOLUME TASKS NEED TOKEN-OPTIMIZED INFRASTRUCTURE TO MAINTAIN PREDICTABLE UNIT COST

SOLUTION 6: WORKLOAD-OPTIMIZED SERVING

| MEMORY FOOTPRINT COMPRESSION



THE SOLUTION

Compress raw weights using quantization (FP8/INT4) and implement dynamic scale-to-zero logic for idle private nodes.



KEY ENABLER

Red Hat AI Inference Server (RHAIS) delivers high-performance serving paired with turnkey **Red Hat Validated Models** (pre-quantized benchmarked for enterprise). **Token FinOps & Usage Attribution.**



BUSINESS IMPACT

Cut hardware memory footprints by up to 50% and achieve predictable infrastructure spend for high-throughput enterprise workloads.

KEY TAKEAWAY: TRADE VARIABLE OPEX FOR FIXED CAPEX TO OPTIMIZE INFRASTRUCTURE ECONOMICS.

PROBLEM #7: CAPTURING MODEL IP

BALANCING APIS & ENTERPRISE ASSETS

Relying exclusively on third-party APIs for routine business logic limits an enterprise from capturing internal model assets



OpEx Balancing

Public cloud APIs provide instant cutting-edge reasoning, but routing 100% of high-volume repetitive domain tasks externally creates continuous operational expenditure.



Building Enterprise IP

Domain fine-tuning and specialized prompt logic tied strictly to vendor-managed endpoints remain locked inside external ecosystems, preventing the enterprise from accumulating portable model IP.

KEY TAKEAWAY: PUBLIC APIS EXCEL AT BROAD CAPABILITY; PRIVATE MODELS ALLOW YOU TO BUILD AND OWN IP.

SOLUTION 7: BUILDING ENTERPRISE MODEL IP

| FROM TOKEN CONSUMER TO ASSET OWNER



THE SOLUTION

Convert internal domain knowledge into portable open-weight model IP using taxonomy-driven synthetic data generation on private enterprise infrastructure.



KEY ENABLER

Red Hat OpenShift AI enables domain experts to fine-tune open-weight models (Granite, Gemma, Llama) directly on private infrastructure without expensive manual data labeling.



BUSINESS IMPACT

Own your fine-tuned model weights as permanent enterprise IP, running high-throughput inference locally at a fixed, predictable compute cost.

KEY TAKEAWAY: COMPLEMENT PUBLIC CLOUD APIS BY CONVERTING DOMAIN KNOWLEDGE INTO ASSETS.

PROBLEM #8: BOUNDARY CONTROLS

| REGULATED WORKLOAD SOVEREIGNTY

Transmitting confidential internal documents, customer PII, or trade secrets across external network boundaries can trigger strict Infosec and compliance roadblocks.



The Boundary Violation

Infosec and Risk Officers freeze production deployments because prompts, RAG context, and vector embeddings must cross public cloud network boundaries to reach external LLM endpoints.



The Regulatory Standstill

Strict compliance frameworks (GDPR, EU AI Act, HIPAA) require local data sovereignty for core regulated workloads, leaving promising pilot applications permanently stranded.

KEY TAKEAWAY: HIGHLY SENSITIVE OR HEAVILY REGULATED WORKLOADS REQUIRE ZERO-TRUST PRIVATE EXECUTION.

SOLUTION 8: GUARANTEEING DATA SOVEREIGNTY

| ZERO-TRUST ON-PREMISE INFERENCE



THE SOLUTION

Run high-throughput inference and vector RAG pipelines 100% air-gapped on private OpenShift clusters behind your corporate firewall.



KEY ENABLER

Red Hat OpenShift AI provides isolated runtime nodes and encrypted vector pipelines, ensuring zero prompt or data leakage to external cloud endpoints.



BUSINESS IMPACT

Satisfy strict Infosec, GDPR, and sovereign compliance requirements while accelerating AI deployment for confidential enterprise use cases.

KEY TAKEAWAY: SECURE & EXTEND PUBLIC CLOUD AI KEEPING SENSITIVE DATA & INFERENCE 100% LOCAL.

PROBLEM #9: HYBRID OPERATIONAL SPLIT

| FRAGMENTED MANAGEMENT & BUDGETS

Managing isolated operational silos and separate procurement contracts for cloud AI and on-premise infrastructure inflates administrative overhead.



Control Plane Fragmentation

Platform teams are forced to maintain disconnected management toolchains—one set of tools for public cloud AI services and a separate stack for private infrastructure.



Procurement Friction

IT leadership struggles to align separate software licensing agreements with existing public cloud committed spend (EDP) agreements, resulting in stranded budget allocations.

KEY TAKEAWAY: DUAL OPERATIONAL SILOS FOR CLOUD AI AND ON-PREM INFRASTRUCTURE DOUBLES YOUR OVERHEAD.

SOLUTION 9: UNIFIED HYBRID OPERATIONS

| CLOUD, ON-PREM & DESKTOP IN ONE CONTROL PLANE



THE SOLUTION

Deploy an OpenShift control plane that bridges developer desktop productivity agents directly to cloud and on-premise clusters via Cloud Marketplaces to burn down public cloud committed spend.



KEY ENABLER

Red Hat AI Enterprise on Cloud Marketplaces allows seamless procurement and unified management spanning local workflows, public cloud compute, bare-metal, and on-premise nodes.



BUSINESS IMPACT

Maximize cloud committed spend value while giving platform teams a single control plane across local developer desktop experimentation, containers, virtual machines, and enterprise AI workloads.

KEY TAKEAWAY: UNIFY YOUR OPERATIONAL CONTROL PLANE FROM LOCAL WORKFLOWS TO HYBRID CLOUD CLUSTERS

PROBLEM #10: THE SKILLS MISMATCH

| THE ROLE & CAPABILITY GAP

AI requires entirely new roles that traditional DevOps and software teams aren't equipped for.

DevOps → MLOps

Transitioning from traditional container hosting to token-optimized load balancing, KV cache management, and continuous AI Eval pipelines.

AppDev → AI/Agentic Dev

Moving from static API integration to non-deterministic orchestration, vector search, MCP tool integration, and prompt engineering.

AI Over-Reliance (Vibe Coding)

Unmonitored developer reliance on AI assistants generates fragile code debt and degrades long-term engineering problem-solving skills.

KEY TAKEAWAY: ORGANIZATIONS LACK THE ROLES REQUIRED TO ADOPT AND STEER COMPREHENSIVE AI INITIATIVES

SOLUTION 10: COLLABORATIVE ENABLEMENT

| PARTNERED CO-ENGINEERING



THE SOLUTION

Deploy Red Hat Consulting and Training to fill operational talent gaps temporarily while actively upskilling teams in parallel.



KEY ENABLER

Hands-on Training Pathways upskill DevOps platform engineering to MLOps, and AppDev teams to AI/Agentic Developers.



BUSINESS IMPACT

Achieve day-one validation success on initial POCs via collaborative partnerships, retaining complete code and architectural ownership.

KEY TAKEAWAY: WE WORK ALONGSIDE YOUR TEAM IN AN ACTIVE PARTNERSHIP UNTIL YOU'RE SELF SUFFICIENT

DELIVERY FAILURES

WHY AI PROJECTS FAIL



Technology

The Trap: Focuses exclusively on model choice, resulting in model instability, hardware silo hoarding, and runtime scaling blockages.



Process

The Trap: Bypassing structural controls for early demo speed, leading to manual script debt and unmonitored alignment drift.



People

The Trap: Forcing ad-hoc architectural shortcuts instead of systematically upskilling teams, resulting in dangerous lock-in.

KEY TAKEAWAY: MOST FAILED AI INITIATIVES ARE NOT TECHNOLOGY FAILURES. THEY ARE DELIVERY FAILURES.

CORE REQUIREMENTS

WHAT TEAMS ACTUALLY NEED



Flexibility

Swap underlying base models cleanly as open-source technology evolves.



Security

Enforce strict access controls and data boundaries over private assets.



Governance

Clear model line-of-custody validation and automated registry policies.



Observability

Real-time visibility into token throughput, bias, and alignment drift.



Portability

Execute identical workflows across on-premise, cloud, or edge environments.



Scalability

Process massive, highly unpredictable concurrent loads efficiently.

KEY TAKEAWAY: TEAMS DON'T NEED RAW TOOLS—THEY NEED STRUCTURAL ARCHITECTURAL RELIABILITY.

THE TAKEAWAY

| PROTOTYPE VS. ENTERPRISE READY

**Stop Building AI Plumbing.
Start Delivering Value.**

Move from testbed experimentation to secure, enterprise-grade production deployment in weeks, not months.

KEY TAKEAWAY: YOU NEED STANDARDIZED INFRASTRUCTURE. ENTERPRISE GOVERNANCE. ABSOLUTE CONTROL.

THE CORE SUMMARY

| THE ENTERPRISE REALITY

//

**The problem is not building AI.
The problem is building AI that survives
contact with the enterprise.**

STANDARDIZED INFRASTRUCTURE • ENTERPRISE GOVERNANCE • ABSOLUTE CONTROL

#	Friction Point	Strategic C-Suite Solution	Primary Business Value
1	API & SDK Volatility	Standard Gateway (MaaS) & Zero-Trust Harness	Decouples app logic; eliminates model lock-in
2	Brittle Data Pipelines	Automated Pipelines & Synthetic Data Generation	Reclaims 80% developer time; democratizes tuning
3	Fragmented GPU Silos	Dynamic Kubernetes Compute Pooling (Kueue)	Boosts GPU utilization from ~20% to 80%+
4	Shadow AI Fleets	Zero-Trust Credentials & Human-in-the-Loop Approval	Neutralizes compliance risk; provides full audit trails
5	Opaque AI Operations	Dedicated Real-Time GenAI Telemetry	Protects SLAs, output safety, and customer experience
6	Workload Economics	Memory Compression (FP8/INT4) & RHAIS Serving	Reduces hardware footprint 50%; stabilizes OpEx
7	Capturing Model IP	Fine-Tuning Open Models on Enterprise Data	Converts OpEx token consumption into owned IP assets
8	Data Sovereignty	100% Air-Gapped On-Premises/Private Inference	Unlocks regulated use cases; prevents PII leakage
9	Hybrid Operational Split	Cloud Marketplace Procurement, Single Control Plane	Burns down cloud commitments (EDPs); simplifies IT
10	The Skills Mismatch	Co-Engineering & Structured Upskilling Pathways	Build long-term internal capabilities without vendor lock-in

APPENDIX

Reference links for all assertions



Appendix: Problem Documentation

Full Unabbreviated URLs

#1: API Volatility

- MCP Spec: <https://modelcontextprotocol.io/>
- Deprecations: <https://platform.openai.com/docs/deprecations>

#2: Brittle Data Plumbing

- Data Science Survey: <https://www.anaconda.com/state-of-data-science>

#3: Fragmented GPU Silos

- K8s Optimization: <https://cast.ai/press-release/2026-state-of-kubernetes-optimization-report/>
- GPU Crisis: <https://www.rtinsights.com/the-gpu-shortage-is-really-a-data-efficiency-crisis/>

#4: Shadow AI Fleets

- OWASP Top 10: <https://checkmarx.com/learn/breaking-down-the-owasp-top-10-for-llm-applications/>
- Samsung Leak: <https://www.letsaskclaire.com/legal/samsung-chatgpt-trade-secret>

#5: Opaque Operations

- vLLM Tuning: <https://docs.vllm.ai/en/stable/configuration/optimization/>
- Air Canada Case: <https://mashable.com/article/air-canada-forced-to-refund-after-chatbot-misinformation>
- DPD Incident: <https://incidentdatabase.ai/cite/631/>

#6: Workload Economics

- vLLM Optimization: <https://developers.redhat.com/articles/2025/10/30/why-vllm-best-choice-ai-inference-today>

#7: Capturing Model IP

- Granite Foundation: <https://www.ibm.com/granite>

#8: Data Sovereignty

- Sovereignty Guide: <https://neuraltrust.ai/blog/data-sovereignty-complete-guide>

Gartner PoC Reality

- 30%+ Abandonment Rate: <https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025>

Executive Summary Reference Matrix

Full Copyable Links

Problem Area	Industry Focus & Benchmark	Primary Documentation Source Full URL
#1 API Volatility	Open Gateway Standards & Protocol Decoupling	https://modelcontextprotocol.io/
#2 Data Plumbing	80% Developer Time Wasted on Non-Value Prep	https://www.anaconda.com/state-of-data-science
#3 GPU Silos	95% GPU Underutilization Due to Static Reservation	https://cast.ai/press-release/2026-state-of-kubernetes-optimization-report/
#4 Shadow AI	Static Keys, Excessive Agency & OWASP Risks	https://checkmarx.com/learn/breaking-down-the-owasp-top-10-for-llm-applications/
#5 Opaque Ops	APM Blind Spots, TTFT & KV Cache Telemetry	https://docs.vllm.ai/en/stable/configuration/optimization/
#6 Workload Costs	Token-Aware Serving Engines vs. CPU Scaling	https://developers.redhat.com/articles/2025/10/30/why-vllm-best-choice-ai-inference-today
#7 Model IP	Fine-Tuned Open Models as Enterprise Assets	https://www.ibm.com/granite
#8 Sovereignty	Zero-Trust Execution & Regulated Workload Bounds	https://neuraltrust.ai/blog/data-sovereignty-complete-guide
Gartner Reality	30%+ GenAI Projects Abandoned Post-PoC	https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025