

Deploying 70B Open-Weight Models in 21 CFR Part 11 & GxP Regulated Facilities

A technical blueprint for running quantized 70B parameter models on on-site hardware while maintaining strict 21 CFR Part 11 audit trails, zero telemetry, and deterministic provenance.

AUTHOR Dr. Carlos Perez Tobie

CATEGORY Architecture & Hardware

PUBLISHED 2026-08-15 / 8 min read

Deploying 70B Open-Weight Models in 21 CFR Part 11 & GxP

When engineering AI systems for pharmaceutical manufacturing, clinical diagnostics, or biotechnology facilities, the standard commercial cloud architecture is a regulatory non-starter. Sending proprietary formulation data, clinical trial patient notes, or unreleased batch records through third-party multi-tenant API endpoints creates immediate compliance liabilities under FDA 21 CFR Part 11, EU Annex 11, and HIPAA data sovereignty mandates.

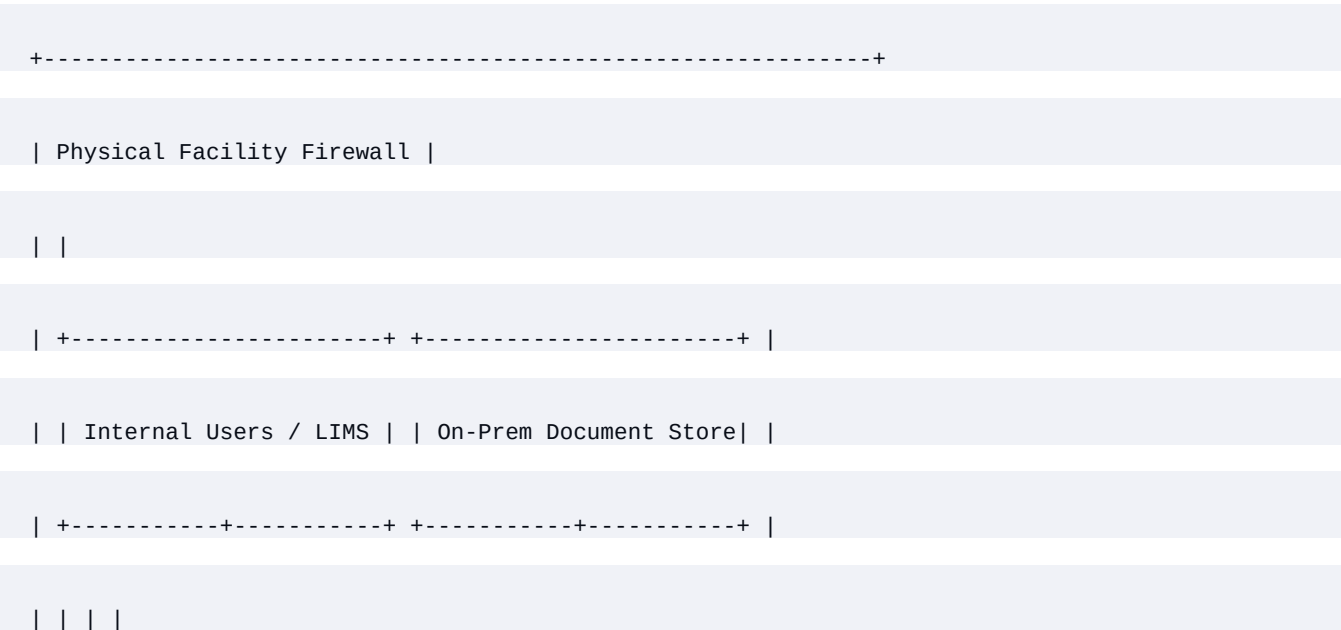
The alternative is not to forgo intelligence, but to shift the execution boundary: run open-weight foundation models (such as Llama 3.3 70B, Qwen 2.5 72B, and DeepSeek variants) on dedicated silicon located entirely inside the client's physical firewall.

Here is how we architect and validate on-premise inference stacks for GxP environments.

1. Physical Isolation and Network Architecture

True air-gapping requires more than disabling outbound ports on a router. In our reference architecture:

- Dual Network Interface Cards (NICs): The inference host operates on two physically segregated network segments. NIC 0 connects solely to an isolated internal storage network hosting document repositories and vector stores. NIC 1 connects to an authenticated internal proxy that serves inference API calls to sanctioned enterprise applications (MES, LIMS, ERP).
- Zero Outbound Gateway: No default gateway is configured on either interface. DNS queries for external domains fail deterministically at the kernel level.
- Telemetry Nullification: Open-source inference runtimes (such as vLLM and TensorRT-LLM) are built from verified source trees with all phone-home telemetry, Hugging Face hub checks, and tracking webhooks compiled out.



```
| v (mTLS Internal) v (Read-Only) |
```

```
| +-----+ |
```

```
| | Isolated Inference Server (Dual NIC) | |
```

```
| | - 4x NVIDIA RTX 6000 Ada / H100 NVL | |
```

```
| | - Open-Weight 70B Parameter Model (FP8) | |
```

```
| | - Append-Only Audit Log Daemon (Part 11) | |
```

```
| +-----+ |
```

```
| |
```

```
| [X] NO OUTBOUND INTERNET ACCESS [X] |
```

```
+-----+
```

2. Model Precision and Hardware Sizing

A 70-billion parameter model in full 16-bit precision requires approximately 140 GB of GPU VRAM simply to load weights, before allocating memory for the KV-cache. In production manufacturing environments where latency and batch concurrency matter, we deploy models using optimized FP8 or INT4 AWQ quantizations:

- FP8 Precision (Ada Lovelace / Hopper): Weights and activations are quantized to 8 bits with less than 0.5% perplexity degradation. A 70B model occupies ~72 GB of VRAM. A dual NVIDIA RTX 6000 Ada (96 GB combined VRAM) or dual H100 NVL setup provides sufficient headroom for a 64k token context window and 16 concurrent requests.
- Latency Targets: Using continuous batching via vLLM, our benchmarked time-to-first-token (TTFT) averages 240ms, with sustained decode throughput exceeding 52 tokens/second per active stream.

3. 21 CFR Part 11 Audit Trail Integrity

Section 11.10 of 21 CFR Part 11 mandates computer systems validation, time-stamped audit trails, and deterministic record attribution. In typical cloud LLM setups, non-deterministic token sampling and opaque vendor updates render validation impossible.

To satisfy GxP validation protocols (IQ/OQ/PQ), our deployments enforce:

- Greedy Sampling for Formal Verification: When agents process regulatory documentation, temperature is locked to 0.0 with deterministic top-p gating. Any query given identical context produces byte-for-byte identical output.
- Cryptographic Model Hash Pinning: At service boot, the SHA-256 checksum of every model tensor file is verified against an immutable validation manifest. If a single byte is altered, the inference service halts.
- Append-Only Local Audit Journal: Every inference request, exact prompt tokens, system prompt version, user credentials, timestamp, and token generation log are written to an append-only cryptographic write-once ledger on a local NVMe array.

4. Operational Handover

A consulting engagement is only successful if the client's internal engineering and compliance teams can operate the system indefinitely without external dependencies. Every Perez Tobie on-premise installation includes:

- Complete infrastructure-as-code manifests (Docker Compose / Podman / Kubernetes)
- Local model artifact repositories mirroring open-weight checkpoints
- Standard operating procedures (SOPs) for model re-quantization and context expansion
- Disaster recovery runbooks for bare-metal hardware restoration within 4 hours

By owning both the physical silicon and the open-weight model weights, regulated enterprises eliminate vendor lock-in, eliminate per-token cloud charges, and build an unassailable data privacy posture.