

Despliegue de Modelos Open-Weight de 70B en Instalaciones Reguladas por 21 CFR Parte 11 y GxP

Un esquema técnico para ejecutar modelos de 70 mil millones de parámetros en hardware local manteniendo pistas de auditoría bajo 21 CFR Parte 11, telemetría cero y procedencia determinista.

AUTHOR Dr. Carlos Perez Tobie

CATEGORY Arquitectura y Hardware

PUBLISHED 2026-08-15 / 8 min de lectura

Despliegue de Modelos Open-Weight de 70B en Instalaciones

Al diseñar sistemas de inteligencia artificial para plantas farmacéuticas, laboratorios de diagnóstico clínico o instalaciones de biotecnología, la arquitectura de nube pública convencional representa un impedimento regulatorio inmediato. Enviar fórmulas propietarias, notas clínicas de pacientes o registros de lote aún no liberados a través de terminales API compartidas de terceros crea contingencias de cumplimiento graves frente a FDA 21 CFR Parte 11, Anexo 11 de la UE y normativas de soberanía de datos en América Latina y Norteamérica.

La solución no consiste en renunciar a la inteligencia artificial, sino en trasladar la frontera de cómputo: operar modelos fundacionales de pesos abiertos (como Llama 3.3 70B, Qwen 2.5 72B y variantes de DeepSeek) en silicio dedicado ubicado estrictamente dentro del perímetro físico del cliente.

A continuación detallamos cómo diseñamos y validamos pilas de inferencia local para entornos GxP.

1. Aislamiento Físico y Arquitectura de Red

El aislamiento por espacio físico (*air-gapping*) requiere mucho más que desactivar puertos en un enrutador. En nuestra arquitectura de referencia:

- Tarjetas de Interfaz de Red Duales (NICs): El servidor de inferencia opera en dos segmentos de red físicamente segregados. La NIC 0 se conecta exclusivamente a una red de almacenamiento interno que hospeda repositorios de documentos y bases de vectores. La NIC 1 se conecta a un proxy interno autenticado que despacha llamadas API hacia aplicaciones corporativas autorizadas (MES, LIMS, ERP).
- Cero Puerta de Enlace Hacia Internet: No se configura ninguna puerta de enlace predeterminada en ninguna interfaz. Las consultas DNS para dominios externos son bloqueadas de forma determinista a nivel de kernel.
- Anulación de Telemetría: Los motores de inferencia de código abierto (como vLLM y TensorRT-LLM) se compilan desde código fuente verificado, eliminando llamadas de telemetría a servidores remotos, verificaciones hacia Hugging Face o rastreadores externos.

2. Precisión del Modelo y Dimensionamiento de Hardware

Un modelo de 70 mil millones de parámetros en precisión completa de 16 bits requiere aproximadamente 140 GB de VRAM en tarjeta gráfica sólo para alojar los pesos en memoria, antes de reservar memoria para la caché de claves y valores (KV-cache). En plantas industriales donde la latencia y la concurrencia son críticas, desplegamos modelos utilizando cuantización optimizada en FP8 o INT4 AWQ:

- Precisión FP8 (Ada Lovelace / Hopper): Los pesos y activaciones se cuantizan a 8 bits con una degradación de perplejidad inferior al 0.5%. El modelo de 70B ocupa ~72 GB de VRAM. Una configuración dual de NVIDIA RTX 6000 Ada (96 GB combinados) o dual H100 NVL ofrece margen suficiente para una ventana de contexto de 64k tokens y 16 solicitudes simultáneas.

- Métricas de Latencia: Utilizando lotes continuos en vLLM, nuestro tiempo al primer token (TTFT) promedia 240 ms, con una tasa sostenida de decodificación superior a 52 tokens/segundo por hilo activo.

3. Integridad de Pista de Auditoría bajo 21 CFR Parte 11

La Sección 11.10 de 21 CFR Parte 11 exige validación de sistemas computarizados, pistas de auditoría con sello temporal y atribución determinista de registros. En modelos en la nube, el muestreo no determinista y las actualizaciones opacas del proveedor anulan la posibilidad de validación formal.

Para satisfacer los protocolos de validación GxP (IQ/OQ/PQ), nuestros despliegues implementan:

- Muestreo Greedy para Verificación Formal: Cuando los agentes procesan documentación regulatoria, la temperatura se bloquea en 0.0. Cualquier consulta con idéntico contexto genera una respuesta byte por byte idéntica.
- Fijación Criptográfica de Hash de Modelo: Al iniciar el servicio, el checksum SHA-256 de cada archivo de tensores se valida contra un manifiesto inmutable. Si un solo byte varía, el servicio de inferencia no inicia.
- Registro Local Inmutable (Append-Only): Cada solicitud de inferencia, tokens exactos de entrada, versión del sistema, usuario, marca de tiempo y log de generación se almacenan en un libro contable local criptográfico en discos NVMe.

4. Transferencia Operativa

Una consultoría sólo tiene éxito si el equipo interno del cliente puede operar el sistema de manera independiente. Cada instalación de Perez Tobie incluye manifiestos de infraestructura como código, manuales de contingencia para recuperación en frío en menos de 4 horas y protocolos para re-cuantización y expansión de contexto.