MATRIZ DE DECISIÓN: CLOUD AI VS. SOVEREIGN AI (EDICIÓN 2026)

Objetivo: Determinar la viabilidad técnica y estratégica de migrar cargas de trabajo de IA a infraestructura propia.

---

1. CRITERIOS DE EVALUACIÓN TÉCNICA

Instrucciones: Evalúa tu caso de uso del 1 al 5 en cada categoría.

Categoría	Cloud AI (APIs Externas)	Sovereign AI (Modelos Locales)
Privacidad del Dato	Compartida con el proveedor. Riesgo de entrenamiento de terceros.	**Total.** El dato nunca sale del perímetro de la empresa.
Latencia	Dependiente de conexión a internet y carga del servidor externo.	**Mínima.** Velocidad de red local e inferencia directa en GPU propia.
Personalización	Limitada a "Prompts" o Fine-tuning caro y restringido.	**Total.** Acceso a pesos del modelo, Fine-tuning y RAG profundo.
Coste Marginal	Pago por token (crece con el uso).	Inversión inicial (CAPEX), pero coste por token cercano a **cero**.
Soberanía	Sujeto a cambios de términos y legislaciones extranjeras.	**Independencia absoluta.** El sistema funciona incluso sin internet.

---

2. EL SEMÁFORO DE DECISIÓN (Roadmap)

🔴 **Cuándo quedarte en la NUBE (Cloud AI):**

- Proyectos en fase de **MVP** que necesitan validación rápida sin inversión en hardware.
- Casos de uso que requieren **conocimiento generalista masivo** que solo modelos tipo GPT-4 pueden ofrecer.
- Baja frecuencia de uso donde el coste de los tokens no justifica la compra de una GPU.

🟢 **Cuándo migrar a SOBERANÍA (Sovereign AI):**

- **Datos Sensibles:** Información legal, médica, financiera o secretos industriales (IP).
- **Uso Intensivo:** Procesos automatizados que generan millones de tokens diarios.
- **Inferencia Crítica:** Sistemas que deben funcionar en tiempo real o en entornos sin conexión garantizada.
- **Especialización:** Cuando necesitas que la IA se comporte como un experto en **tus** procesos específicos de ingeniería.

---

3. CHECKLIST DE INFRAESTRUCTURA MÍNIMA

Si decides ir por el camino de la **IA Soberana**, asegúrate de cumplir estos requisitos:

- [ ] **Hardware:** Mínimo una GPU con 24GB de VRAM (ej. NVIDIA RTX 4090 o A6000) para modelos de 7B/14B.
- [ ] **Software:** Implementación de servidores de inferencia como *Ollama*, *vLLM* o *LocalAI*.
- [ ] **Base de Datos Vectorial:** Instalación de *ChromaDB* o *Pinecone* (versión local) para el sistema RAG.
- [ ] **Talento:** Capacidad interna para gestionar contenedores (Docker/Kubernetes).