Procedimiento de Gestión del Ciclo de Vida de un LLM de Alto Riesgo

Empresa: CibersegurIA S.L. | Sistema: CibersegurIA-Licita | Versión: 1.0 | Fecha: Septiembre 2025

Este procedimiento establece cómo se planifica, opera, monitoriza y mejora el sistema LLM "CibersegurIA-Licita" para asegurar su cumplimiento continuo con el AI Act (Art. 9, 10, 11, 14 y Cap. IV), ENS (nivel ALTO), RGPD y normativa sectorial. Incluye roles, controles, evidencias y criterios de aceptación.

1. Objeto y alcance

• Objeto: Garantizar que el LLM mantiene requisitos de seguridad, trazabilidad, exactitud, explicabilidad y supervisión humana durante todo su ciclo de vida.
• Alcance: Desde diseño y entrenamiento hasta operación, retrain, cambios sustanciales y fin de vida.
• Entornos: Desarrollo, preproducción (UAT), producción.
• Normativa aplicable: AI Act (Art. 6, 9–16, 43; Anexo IV), RGPD (Art. 5, 25, 32, 35), ENS (MP/OP/CP), ENI.

2. Definiciones clave

• Cambio sustancial: Modificación que afecta a comportamiento del modelo (p. ej., nuevo dataset >5% del corpus, alteración de hiperparámetros principales, cambio de arquitectura RAG o proveedor vectorial, nuevas funciones que influyen en decisiones).
• Retrain: Reentrenamiento del modelo con datos nuevos o reajustes significativos.
• Evidencias: Artefactos verificables (logs firmados, informes de pruebas, actas del Comité de IA).

3. Gobierno y RACI

Actividad	IA Officer	DPO/DPD	CISO/IT Ops	Owner de Negocio
Aprobación de cambios sustanciales	R/A	C	C	A
Validación pre-despliegue	A	C	R	C

Gestión de incidencias	C	C	R/A	C
Supervisión humana / Uso responsable	A	C	C	R
Reporte a autoridades / clientes	A	R	C	C
Auditorías internas	A	C	R	C

Leyenda: R = Responsable (ejecuta) | A = Aprueba | C = Consultado

4. Control de cambios y versionado

• Registro CC: Cada modificación se documenta en el "Registro de Cambios del LLM" con ID, fecha, autor, descripción, riesgo, pruebas asociadas y decisión (go/no-go).
• Cambio sustancial → exige revalidación completa y actualización de documentación técnica.
• Ejemplo real: CC-2025-014: Actualización del corpus con 1.240 pliegos (2024Q4). Aprobación Comité IA (Acta 2025-11). Ruta evidencias: \\Repositorio\LLM\Cambios\2025\CC-2025-014\

5. Política de datos y retrain

• Frecuencia estándar: Retrain anual (noviembre) o inducido por evento (p. ej., nuevas guías regulatorias, cambio de distribución del dominio, >5% de documentos nuevos).
• Dataset delta: Documentar fuentes, limpieza, exclusiones, anonimización/pseudonimización. Ej.: \\Datasets\Licita\Delta_2024Q4\Informe_Curación.pdf
• Lista de bloqueo: exclusión de documentos obsoletos o no representativos (ver \\Datasets\Licita\Listas\Blacklist_v3.csv).

6. Validación previa a despliegue (go/no-go)

• Pruebas funcionales: 300 Q&A representativas (sampling estratificado por tipo de cláusula). Criterio: precisión ≥90%, cobertura ≥85%.
• Trazabilidad: 100% de respuestas deben citar documento y fragmento (offset) o marcar "insuficiente evidencia".
• Sesgo y robustez: pruebas con prompts adversariales y variantes lingüísticas; desviación máxima tolerable ±5%.
• Seguridad: pentest sobre endpoints críticos, revisión de permisos RBAC, secretos rotados y escaneo SCA.
• Actas de validación: \\Calidad\LLM\Validaciones\2025\VAL-2025-03\Informe_Final.pdf

7. Monitorización continua en producción

• KPIs base (mensuales):
  – % respuestas con fuente citada (objetivo ≥95%).
  – % respuestas con semáforo de confianza “verde” (≥80%).
  – % escalados a revisión humana (<15%).
  – TTR de incidencias de contenido (≤72 h) y de seguridad (≤24 h críticas).
  – Drift de distribución vs. última validación (<10%).
• Alertas automáticas: desviaciones >X% generan ticket y revisión del Comité de IA.

8. Gestión de incidencias y problemas

• Tipología: (a) Contenido incorrecto/obsoleto; (b) Sesgo; (c) Alucinación; (d) Seguridad; (e) Privacidad.
• Canal: Formulario “INC-LLM” en Service Desk. Evidencias mínimas: prompt, respuesta, hash de índice, usuario, timestamp.
• SLA: Clasificación en 8 h; contención en 24 h (críticas); RCA en 5 días hábiles.
• Cierre: Acción correctiva/preventiva, actualización de riesgo y de dataset si procede. Ruta: \\Incidencias\LLM\2025\INC-XXXX\

9. Seguridad (ENS) y control de acceso

• ENS ALTO: MP-01 (protección en tránsito TLS1.3), MP-05 (cifrado en reposo AES-256), CP-10 (control de acceso lógico y SoD), OP-04 (gestión de vulnerabilidades: críticas <24h), MP-08 (trazabilidad con logs firmados y clock sync).
• Acceso por rol (RBAC): analista, revisor, owner, auditor. MFA obligatorio. Revisiones trimestrales de permisos.
• Pruebas de seguridad: semestrales o tras cambios sustanciales. Evidencias: \\Seguridad\Pentest\LLM\2025\

10. Supervisión humana efectiva (Art. 14 AI Act)

• Ninguna salida del LLM es ejecutiva por sí sola. La interfaz obliga a marcar “Revisión humana realizada”.
• Perfil revisor: personal experto en contratación pública o jurídico-administrativo.
• Muestreo aleatorio mensual del 5% de respuestas “verdes” para auditoría interna.

11. Documentación, registros y conservación

• Dossier por versión: dataset (hashes), pruebas, actas, checklist de cumplimiento, manuales y changelog.
• Conservación mínima: 10 años tras fin de uso (Art. 11.1.c AI Act). Copias en bóveda WORM. Rutas: \\Conformidad\AIAct\CibersegurIA-Licita\Versiones\v1_0\

12. Gestión de terceros y dependencias

• Due diligence: licencias, subprocesadores, soporte, SLA de seguridad (parches críticos <24h), ubicación de datos UE.
• Cambios de proveedor/modelo base → evaluación de impacto y, si aplica, nueva conformidad.
• Evidencias: \\Terceros\Evaluaciones\2025\DD-LLM-VectorDB.pdf

13. Retirada controlada (EoL) y rollback

• Criterios: obsolescencia técnica/regulatoria, riesgos no mitigables, baja relación coste-beneficio.
• Plan: comunicación a usuarios, congelación de índices, export de evidencias, borrado seguro (NIST 800-88), revocación de credenciales y cierre de endpoints. Informe EoL: \\Conformidad\AIAct\EoL\EOL-2027-01.pdf

14. Anexos operativos

A) Plantilla de Registro de Cambios (extracto):

ID Cambio	Descripción	Tipo	Riesgo	Evidencias	Decisión
CC-2025-014	Añadido corpus 2024Q4 (1.240 pliegos)	Sustancial	Medio–Alto	\\Repositorio\LLM\Cambios\2025\CC-2025-014*	GO (Acta 2025-11)

B) Checklist de despliegue (pre-producción → producción):

□ Pruebas funcionales ≥90% precisión ▢ OK
□ Cobertura ≥85% (muestras representativas) ▢ OK
□ Fuentes citadas en 100% de respuestas ▢ OK
□ Pentest y SCA sin hallazgos críticos ▢ OK
□ Revisión DPO/DPIA actualizada ▢ OK
□ Comité IA aprueba (Acta) ▢ OK

C) KPIs y umbrales (tablero mensual):

KPI	Objetivo	Umbral alerta	Fuente
% respuestas con fuente citada	≥95%	<90%	Logs chatbot (export CSV)
% confianza “verde”	≥80%	<70%	Panel explicabilidad

% escalados a revisión	≤15%	>20%	Registro humano
TTR incidencias contenido	≤72h	>96h	Service Desk
Vulns críticas cerradas	100% <24h	<100%	SIEM / Gestor parches

D) Plantilla de incidente (INC-LLM):

• Datos mínimos: ID, fecha, usuario, prompt, respuesta, doc fuente/offset, hash índice, severidad, impacto, acciones, RCA, evidencias.
• Ejemplo ruta: \\Incidencias\LLM\2025\INC-2025-073\RCA.pdf