

ISLA — Isolated System Legal Assistant

Un asistente jurídico con IA multimodal, 100 % local y aislado, para portátiles M1

Ricardo Sánchez — Nesilabs Tech S.L.

España · Informe técnico

INFORME TÉCNICO

RESUMEN

Presentamos **ISLA** (Isolated System Legal Assistant), una aplicación de escritorio para portátiles M1 que ejecuta un asistente jurídico con IA de forma completamente local y sin conexión. ISLA ingiere documentos legales heterogéneos (PDF nativo y escaneado, Word, imágenes y audio), los convierte en contexto consultable mediante OCR, visión y transcripción, y responde en español sobre el material del usuario con citas exactas a nivel de archivo, página y línea, así como a norma y artículo sobre una biblioteca de legislación española preindexada. El diseño se articula en torno a tres niveles de aislamiento —red (air-gapped), equipo (cada Mac es una isla) y datos en reposo (cifrado ligado al Llavero)— que convierten la confidencialidad del secreto profesional en una propiedad técnica y no en una promesa contractual. Describimos la arquitectura del sistema, el pipeline de recuperación híbrida (BM25 + vectorial con fusión RRF y reranking), el modelo de seguridad y licenciamiento por hardware (Ed25519), y las razones por las que la memoria unificada y la aceleración Metal del chip M1 hacen viable ejecutar un modelo multimodal en un portátil.

1. Introducción

Motivación. Los despachos y departamentos jurídicos manejan los datos más sensibles que existen: datos personales de clientes, estrategia procesal, prueba y secreto profesional. Las herramientas de IA generalistas basadas en la nube envían ese contenido a servidores de terceros, lo que puede entrar en conflicto con el deber de confidencialidad, el RGPD/LOPDGDD y el propio secreto profesional, y expone el material a caché, indexación o reutilización para entrenamiento ajeno. En este sector, la nube no es una comodidad sino un riesgo estructural.

Propuesta. ISLA lleva la capacidad de un asistente conversacional jurídico al interior del propio equipo. Toda la inferencia, la lectura de documentos, la búsqueda y la generación ocurren en local. El principio de diseño es simple: los datos entran, se procesan y permanecen en la máquina.

Contribuciones.

1. Una arquitectura de asistente jurídico multimodal que funciona íntegramente offline en un único binario autocontenido.
2. Un modelo de aislamiento en tres niveles (red, equipo, reposo) que hace de la confidencialidad una

imposibilidad técnica de exfiltración, no una cláusula.

3. Un pipeline de recuperación híbrida con citación verificable a nivel de página/línea y de norma/artículo sobre legislación española preindexada.

2. Modelo de amenaza y objetivos de diseño

El adversario relevante no es un atacante remoto sofisticado, sino la *superficie de fuga por diseño*: cualquier ruta por la que un fragmento de expediente pueda abandonar el equipo. El objetivo es eliminar esa superficie. De ahí tres invariantes:

- **Sin salida de red.** El funcionamiento no requiere ninguna llamada a internet; puede operar con el Wi-Fi apagado o en red aislada.
- **Sin punto único de fallo.** No hay servidor central ni cuenta en la nube; comprometer un caso exigiría acceso físico al Mac concreto y a la sesión del usuario.
- **Ilegible en reposo.** El contenido y el índice de búsqueda se guardan cifrados; la clave vive en el Llavero, encadenada al login y a FileVault.

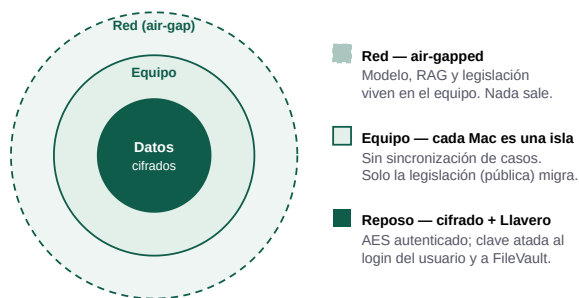


Figura 1. Los tres niveles de aislamiento. La confidencialidad se obtiene por construcción: no hay a dónde viajen los datos, no hay servidor que comprometer y, sin la sesión del usuario, el material en disco es ilegible.

3. Arquitectura del sistema

ISLA se distribuye como una única aplicación nativa (~15 GB) que empaqueta el intérprete, los modelos y el motor de inferencia, de modo que se instala sin dependencias, sin internet y sin conocimientos técnicos. En ejecución consta de tres piezas locales que dialogan sobre `127.0.0.1`: una interfaz React/Tailwind en una ventana nativa (WebView), un backend FastAPI ligero (un proceso, un puerto) y el motor de inferencia que sirve el modelo. El motor arranca de forma automática en segundo plano y precarga el modelo, de modo que la app está lista al instante al abrirse.

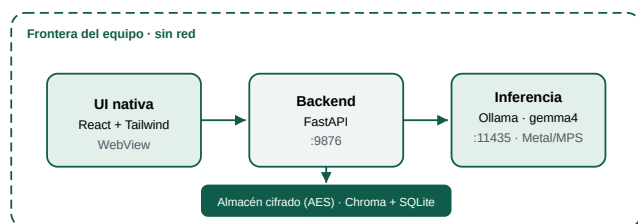


Figura 2. Arquitectura en ejecución. Tres componentes locales sobre `localhost`; todo el estado sensible se persiste cifrado. Ningún puerto se expone fuera del equipo.

4. Pipeline de recuperación (RAG)

Las respuestas se anclan en las fuentes del usuario mediante recuperación aumentada. Cada documento se trocea (para legislación, *por artículo*) y se indexa por dos vías complementarias: un índice vectorial (ChromaDB) sobre *embeddings* de significado y un índice léxico BM25

sobre palabras clave. En consulta, ambos rankings se combinan mediante fusión recíproca de rangos (RRF) y un *cross-encoder* reordena los candidatos para que la cita correcta aparezca primero. El modelo genera entonces la respuesta citando archivo, página y línea —y norma y artículo cuando la fuente es legislación.

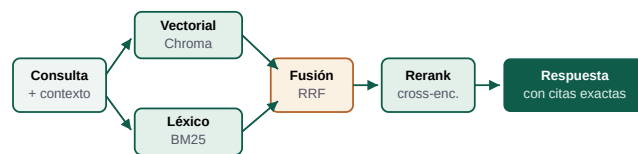


Figura 3. Recuperación híbrida. El índice semántico y el léxico se fusionan con RRF y se reordenan con un *cross-encoder*; el modelo responde citando la fuente exacta, lo que permite abrir y verificar cada afirmación.

5. Ingesta multimodal

ISLA convierte material que no es texto en contexto consultable. Los PDF nativos y de Word se extraen directamente; los escaneados y las fotos pasan por OCR (Tesseract, español e inglés); los diagramas, tablas, sellos y manuscritos se interpretan con el componente de visión del modelo; y el audio —declaraciones, vistas, notas de voz— se transcribe con Whisper large-v3 acelerado en Apple Silicon (MLX). Todo desemboca en el mismo almacén cifrado e indexado.

6. Base de conocimiento jurídico

ISLA incluye una biblioteca de legislación española preconstruida dentro de la aplicación: más de 30 códigos y leyes (Constitución, Código Civil, Código Penal, LEC, LECrim, Estatuto de los Trabajadores, Ley Concursal, LSC, LOPJ, entre otros) con aproximadamente 22.000 artículos, troceados *por artículo* para permitir citación precisa. El índice viaja ya construido y se copia a la cuenta la primera vez que arranca, de modo que la biblioteca está disponible sin descargar nada de internet. La legislación es contenido público y es lo único que puede transferirse entre equipos; el material de cada caso nunca se sincroniza.

7. Modelo de seguridad y licenciamiento

La seguridad de ISLA se encadena a la de la plataforma. El contenido sensible se cifra en reposo con AES autenticado y la clave se custodia en el Llavero de macOS, protegido por el login del usuario, Touch ID y FileVault; los

originales no se dispersan por el disco: tras procesarlos solo queda el conocimiento cifrado y sus metadatos. El acceso puede reforzarse con doble factor (TOTP), que bloquea la app tras reinicio, cierre de tapa o suspensión.

La protección del propio software es independiente de la de los datos. Cada instalación se activa con una licencia firmada criptográficamente con **Ed25519** y atada a la huella de hardware del Mac concreto; la aplicación solo incorpora la clave pública para *verificar* la firma, de modo que una licencia no es transferible a otro equipo. El código propietario se distribuye compilado a binario nativo, no como fuente interpretable.

8. Despliegue en portátiles M1

La elección del chip M1 como único objetivo es deliberada, y la razón central es la **memoria unificada**. A diferencia de un PC convencional, donde la GPU tiene su propia VRAM separada de la RAM del sistema, en el M1 la CPU y la GPU comparten un mismo espacio físico de memoria. Eso significa que el modelo de IA se carga una sola vez en esa memoria común y la GPU accede a él directamente, sin copiarlo de la RAM a una tarjeta gráfica. En la práctica permite ejecutar con fluidez un modelo multimodal de varios gigas en un portátil —un equipo con 16–32 GB lo mueve con holgura— algo que en un PC exigiría una tarjeta gráfica dedicada, cara y voluminosa, con VRAM suficiente.

Por qué la gama Max, no la Pro. Dentro de la misma familia M1, la variante Max no destaca tanto por más núcleos como por el **ancho de banda de memoria**: es aproximadamente el doble que el de la variante Pro. En inferencia local ese ancho de banda es el cuello de botella real —generar cada token exige recorrer los pesos del modelo en memoria—, de modo que más ancho de banda se traduce directamente en más velocidad de respuesta. Para ISLA, un M1 Max rinde de forma notablemente más ágil que un M1 Pro con el mismo modelo, precisamente por esa diferencia de ancho de banda.

A esto se suma la **aceleración Metal/MPS y el Neural Engine**, que aceleran la inferencia, la transcripción

9. Discusión, limitaciones y conclusión

Compromisos. Operar 100 % en local fija un techo de tamaño de modelo al hardware del usuario: ISLA prioriza un modelo de 8B que corre con soltura en un portátil frente a un modelo mayor inalcanzable sin nube. Es un intercambio consciente de capacidad bruta por confidencialidad absoluta y coste de pago único, sin suscripción de tokens.



Figura 4. Cadena de confianza. Cada capa se apoya en la inferior; el cifrado de la aplicación se ancla en la seguridad de plataforma del propio Mac, y la licencia queda ligada al identificador de hardware.

(Whisper MLX) y el reranking de serie y sin configuración; la **seguridad de plataforma** (FileVault, Secure Enclave, Gatekeeper, Llavero), que aporta la cadena de confianza sobre la que se apoya el cifrado de la app; y el hecho de centrarse en un objetivo de hardware coherente, que elimina la lotería de *drivers* y configuraciones del ecosistema PC.

Requisitos. Portátil M1 y macOS 14 (Sonoma) o superior. Se recomienda 16–32 GB de RAM (32 GB para máxima holgura) y, para inferencia más ágil, la variante Max por su mayor ancho de banda de memoria.

Componentes del sistema

Capa	Tecnología
Modelo (multimodal)	gemma4 8B vía Ollama, en local
Recuperación	ChromaDB (vectorial) + BM25, fusión RRF
Embeddings	mxbai-embed-large
Reranking	cross-encoder bge-reranker-v2-m3 (MPS)
OCR	Tesseract (es + en)
Visión	gemma4 vision
Audio	Whisper large-v3 (MLX)
Cifrado	AES/Fernet + Llavero de macOS
Acceso	2FA TOTP
Licencia	firma Ed25519 atada al hardware
Backend / UI	FastAPI · React+Tailwind (WebView)

Limitaciones. (i) La calidad de respuesta está acotada por la del modelo local y por la cobertura del corpus del usuario; (ii) la biblioteca legislativa refleja el estado en el momento de empaquetado y, salvo regeneración explícita desde fuentes abiertas oficiales, no incorpora cambios normativos posteriores; (iii) el aislamiento por equipo, que es una fortaleza de seguridad, implica que no hay

colaboración multi-dispositivo sobre un mismo caso por diseño; (iv) como todo sistema asistido por IA, las respuestas requieren verificación profesional —de ahí el énfasis en citas abribles y comprobables.

Conclusión. ISLA muestra que un asistente jurídico multimodal de utilidad práctica puede ejecutarse por completo dentro de un portátil, sin que ningún dato abandone el equipo. Al convertir la confidencialidad en una propiedad estructural —imposibilidad técnica de

exfiltración— en lugar de una cláusula de privacidad, ISLA encaja las obligaciones de secreto profesional y protección de datos del sector jurídico con la potencia de un asistente conversacional moderno, sobre la base de hardware y seguridad de plataforma del portátil M1.

Documento técnico. La descripción se centra en arquitectura y diseño; no incluye una evaluación cuantitativa formal, que queda como trabajo futuro (calidad de recuperación y de respuesta sobre corpus jurídicos de referencia).

Apéndice A. Interfaz

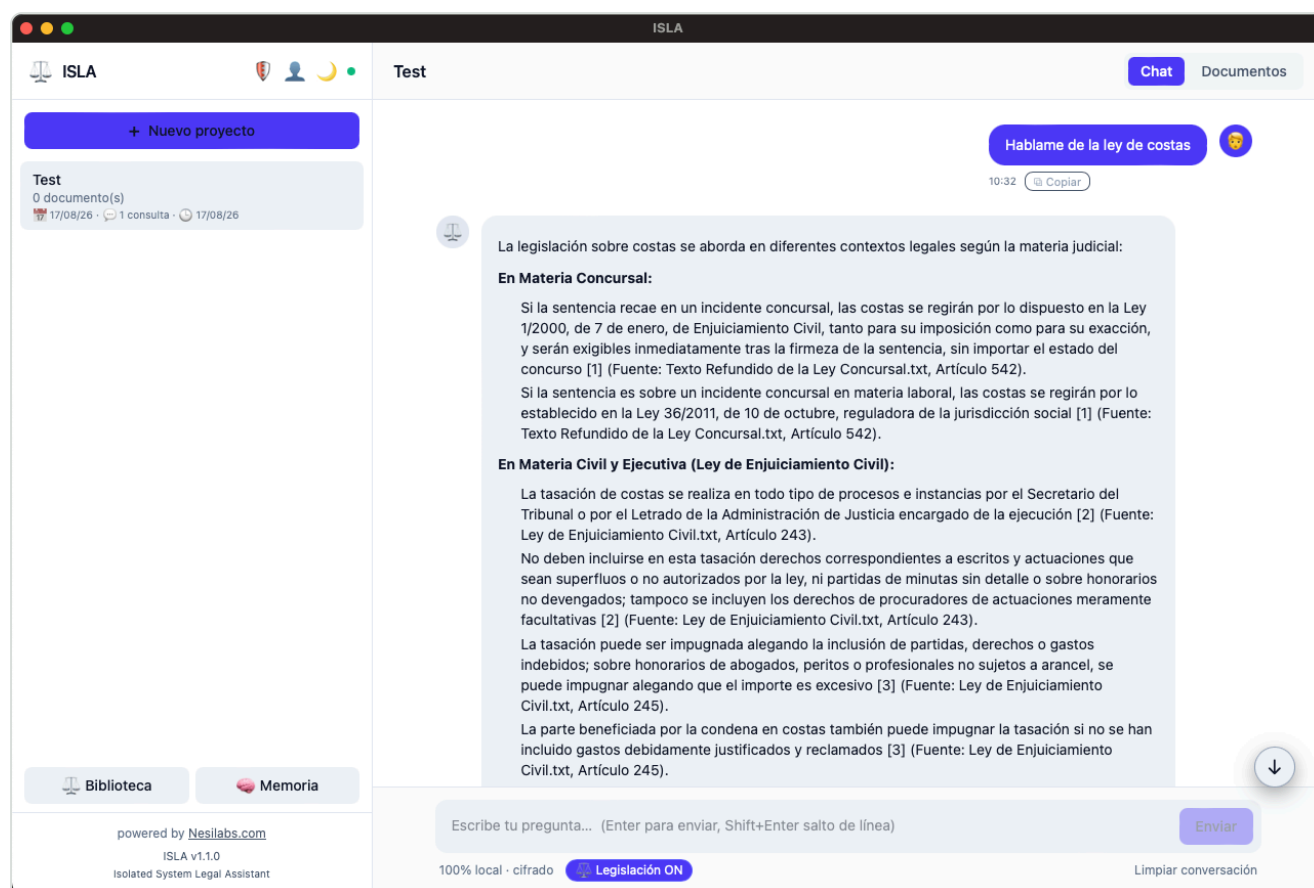


Figura 5. Vista de chat. Respuesta sobre la ley de costas anclada en la legislación, con cita a norma y artículo (p. ej. Ley de Enjuiciamiento Civil, art. 243/245) y marcadores [n] por fuente. Barra inferior: 100 % local · cifrado y el conmutador Legislación ON.

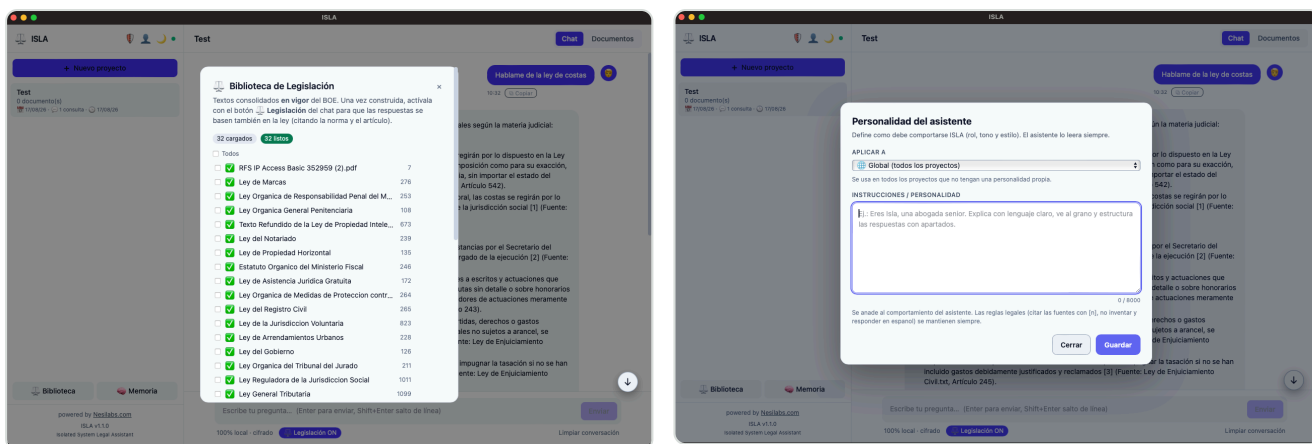


Figura 6. A la izquierda, la biblioteca de legislación preindexada (textos consolidados en vigor del BOE, con recuento de artículos por norma). A la derecha, la personalidad del asistente: instrucciones de rol/tono aplicables global o por proyecto, manteniéndose siempre las reglas de citar fuentes y responder en español.