

ISLA — Isolated System Legal Assistant

A fully local, air-gapped multimodal legal AI assistant for M1 laptops

Ricardo Sánchez — Nesilabs Tech S.L.

Spain · Technical report

TECHNICAL REPORT

ABSTRACT

We present **ISLA** (Isolated System Legal Assistant), a desktop application for M1 laptops that runs a legal AI assistant fully on-device and offline. ISLA ingests heterogeneous legal material (native and scanned PDF, Word, images and audio), turns it into queryable context via OCR, vision and transcription, and answers in Spanish over the user's own material with exact citations down to file, page and line — as well as to statute and article over a pre-indexed Spanish legislation library. The design is organised around three levels of isolation — network (air-gapped), device (each Mac is an island) and data at rest (encryption bound to the Keychain) — which turn attorney confidentiality into a technical property rather than a contractual promise. We describe the system architecture, the hybrid retrieval pipeline (BM25 + vector with RRF fusion and reranking), the security and hardware-bound licensing model (Ed25519), and why the unified memory and Metal acceleration of the M1 make it feasible to run a multimodal model on a laptop.

1. Introduction

Motivation. Law firms and in-house legal teams handle the most sensitive data that exists: clients' personal data, litigation strategy, evidence and professional secrecy. Cloud-based general-purpose AI tools send that content to third-party servers, which can conflict with the duty of confidentiality, the GDPR/LOPDGDD and attorney–client privilege itself, and exposes the material to caching, indexing or reuse for third-party training. In this sector, the cloud is not a convenience but a structural risk.

Proposal. ISLA brings the capability of a conversational legal assistant inside the device itself. All inference, document reading, search and generation happen locally. The design principle is simple: data comes in, is processed, and stays on the machine.

Contributions.

1. A multimodal legal-assistant architecture that runs entirely offline in a single self-contained binary.
2. A three-level isolation model (network, device, at rest) that makes confidentiality a technical impossibility of exfiltration, not a clause.
3. A hybrid retrieval pipeline with verifiable citation at page/line and statute/article level over pre-indexed Spanish legislation.

2. Threat model and design goals

The relevant adversary is not a sophisticated remote attacker but the *leakage-by-design surface*: any route through which a fragment of a case file could leave the device. The goal is to eliminate that surface. Hence three invariants:

- **No network egress.** Operation requires no internet call whatsoever; it can run with Wi-Fi off or on an isolated network.
- **No single point of failure.** There is no central server or cloud account; compromising a case would require physical access to the specific Mac and the user's session.
- **Unreadable at rest.** Content and the search index are stored encrypted; the key lives in the Keychain, chained to the login and to FileVault.

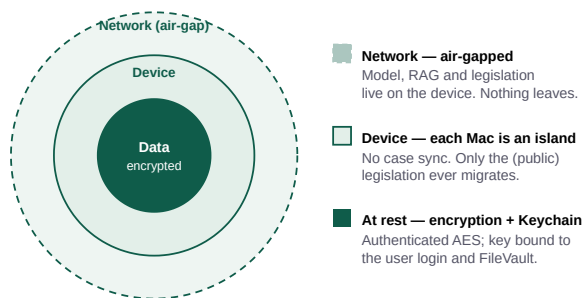


Figure 1. The three levels of isolation. Confidentiality is obtained by construction: there is nowhere for the data to travel, no server to compromise, and without the user's session the material on disk is unreadable.

3. System architecture

ISLA ships as a single native application (~15 GB) bundling the interpreter, the models and the inference engine, so it installs with no dependencies, no internet and no technical knowledge. At runtime it consists of three local pieces that talk over `127.0.0.1`: a React/Tailwind UI in a native window (WebView), a lightweight FastAPI backend (one process, one port) and the inference engine serving the model. The engine starts automatically in the background and preloads the model, so the app is ready instantly on open.

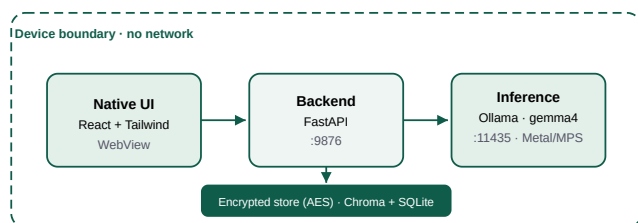


Figure 2. Runtime architecture. Three local components over `localhost`; all sensitive state is persisted encrypted. No port is exposed off the device.

4. Retrieval pipeline (RAG)

Answers are grounded in the user's sources through retrieval-augmented generation. Each document is chunked (for legislation, *per article*) and indexed two complementary ways: a vector index (ChromaDB) over

7. Security and licensing model

ISLA's security is chained to the platform's. Sensitive content is encrypted at rest with authenticated AES and the key is held in the macOS Keychain, protected by the user login, Touch ID and FileVault; originals are not scattered across the disk: after processing, only the encrypted

meaning *embeddings* and a BM25 lexical index over keywords. At query time both rankings are combined via reciprocal rank fusion (RRF) and a *cross-encoder* reorders the candidates so the correct citation surfaces first. The model then generates the answer citing file, page and line — and statute and article when the source is legislation.

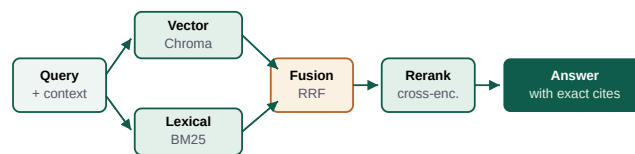


Figure 3. Hybrid retrieval. The semantic and lexical indexes are fused with RRF and reordered with a *cross-encoder*; the model answers citing the exact source, so every claim can be opened and verified.

5. Multimodal ingestion

ISLA turns non-text material into queryable context. Native and Word PDFs are extracted directly; scans and photos go through OCR (Tesseract, Spanish and English); diagrams, tables, stamps and handwriting are interpreted with the model's vision component; and audio — testimony, hearings, voice notes — is transcribed with Whisper large-v3 accelerated on the M1 (MLX). Everything lands in the same encrypted, indexed store.

6. Legal knowledge base

ISLA includes a Spanish legislation library pre-built inside the application: more than 30 codes and statutes (the Constitution, Civil Code, Criminal Code, Civil Procedure Act, Criminal Procedure Act, Workers' Statute, Insolvency Act, Companies Act, Judiciary Act, among others) with roughly 22,000 articles, chunked *per article* to allow precise citation. The index travels pre-built and is copied to the account the first time it starts, so the library is available without downloading anything from the internet. Legislation is public content and is the only thing that can be transferred between devices; each case's material is never synced.

knowledge and its metadata remain. Access can be reinforced with two-factor authentication (TOTP), which locks the app after restart, lid close or sleep.

Protecting the software itself is independent of protecting the data. Each installation is activated with a license cryptographically signed with **Ed25519** and bound to the specific Mac's hardware fingerprint; the application only carries the public key to *verify* the signature, so a license is not transferable to another device. Proprietary code is distributed compiled to a native binary, not as interpretable source.

8. Deployment on M1 laptops

Targeting the M1 chip alone is deliberate, and the central reason is **unified memory**. Unlike a conventional PC, where the GPU has its own VRAM separate from system RAM, on the M1 the CPU and GPU share one physical memory space. This means the AI model is loaded once into that common memory and the GPU accesses it directly, without copying it from RAM to a graphics card. In practice this lets a multi-gigabyte multimodal model run smoothly on a laptop — a machine with 16–32 GB handles it comfortably — something a PC would require a dedicated, expensive and bulky graphics card with enough VRAM to do.

Why the Max tier, not Pro. Within the same M1 family, the Max variant stands out not so much for more cores as for **memory bandwidth**: roughly double that of the Pro variant. In local inference that bandwidth is the real bottleneck — generating each token requires traversing the model's weights in memory — so more bandwidth translates directly into faster responses. For ISLA an M1 Max runs markedly more responsively than an M1 Pro with the same model, precisely because of that bandwidth difference.

On top of this, **Metal/MPS acceleration and the Neural Engine** speed up inference, transcription (Whisper MLX) and reranking out of the box and without configuration;

9. Discussion, limitations and conclusion

Trade-offs. Running 100 % locally caps model size at the user's hardware: ISLA prioritises an 8B model that runs smoothly on a laptop over a larger one unreachable without the cloud. It is a conscious trade of raw capability for absolute confidentiality and a one-off cost, with no token subscription.

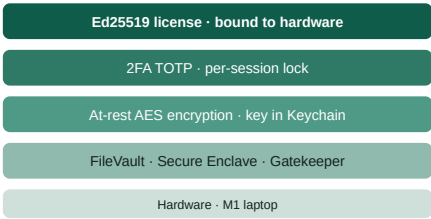


Figure 4. Chain of trust. Each layer rests on the one below; the app's encryption is anchored in the Mac's own platform security, and the license is bound to the hardware identifier.

platform security (FileVault, Secure Enclave, Gatekeeper, Keychain) provides the chain of trust on which the app's encryption rests; and focusing on a coherent hardware target removes the lottery of drivers and configurations of the PC ecosystem.

Requirements. M1 laptop and macOS 14 (Sonoma) or higher. 16–32 GB of RAM is recommended (32 GB for maximum headroom) and, for snappier inference, the Max variant for its higher memory bandwidth.

System components

Layer	Technology
Model (multimodal)	gemma4 8B via Ollama, local
Retrieval	ChromaDB (vector) + BM25, RRF fusion
Embeddings	mxlbai-embed-large
Reranking	cross-encoder bge-reranker-v2-m3 (MPS)
OCR	Tesseract (es + en)
Vision	gemma4 vision
Audio	Whisper large-v3 (MLX)
Encryption	AES/Fernet + macOS Keychain
Access	2FA TOTP
License	Ed25519 signature bound to hardware
Backend / UI	FastAPI · React+Tailwind (WebView)

Limitations. (i) Answer quality is bounded by that of the local model and by the coverage of the user's corpus; (ii) the legislation library reflects its state at packaging time and, unless explicitly regenerated from official open sources, does not incorporate later legal changes; (iii) per-device isolation, a security strength, means there is no multi-device collaboration over the same case by design; (iv) as with any AI-assisted system, answers require

professional verification — hence the emphasis on openable, checkable citations.

Conclusion. ISLA shows that a practically useful multimodal legal assistant can run entirely inside a laptop, with no data leaving the device. By making confidentiality a structural property — a technical impossibility of exfiltration — rather than a privacy clause, ISLA

reconciles the professional-secrecy and data-protection duties of the legal sector with the power of a modern conversational assistant, on the basis of the M1 laptop's hardware and platform security.

Technical document. The description focuses on architecture and design; it does not include a formal quantitative evaluation, which is left as future work (retrieval and answer quality over reference legal corpora).

Appendix A. Interface

Screenshots show the shipping Spanish-language UI.

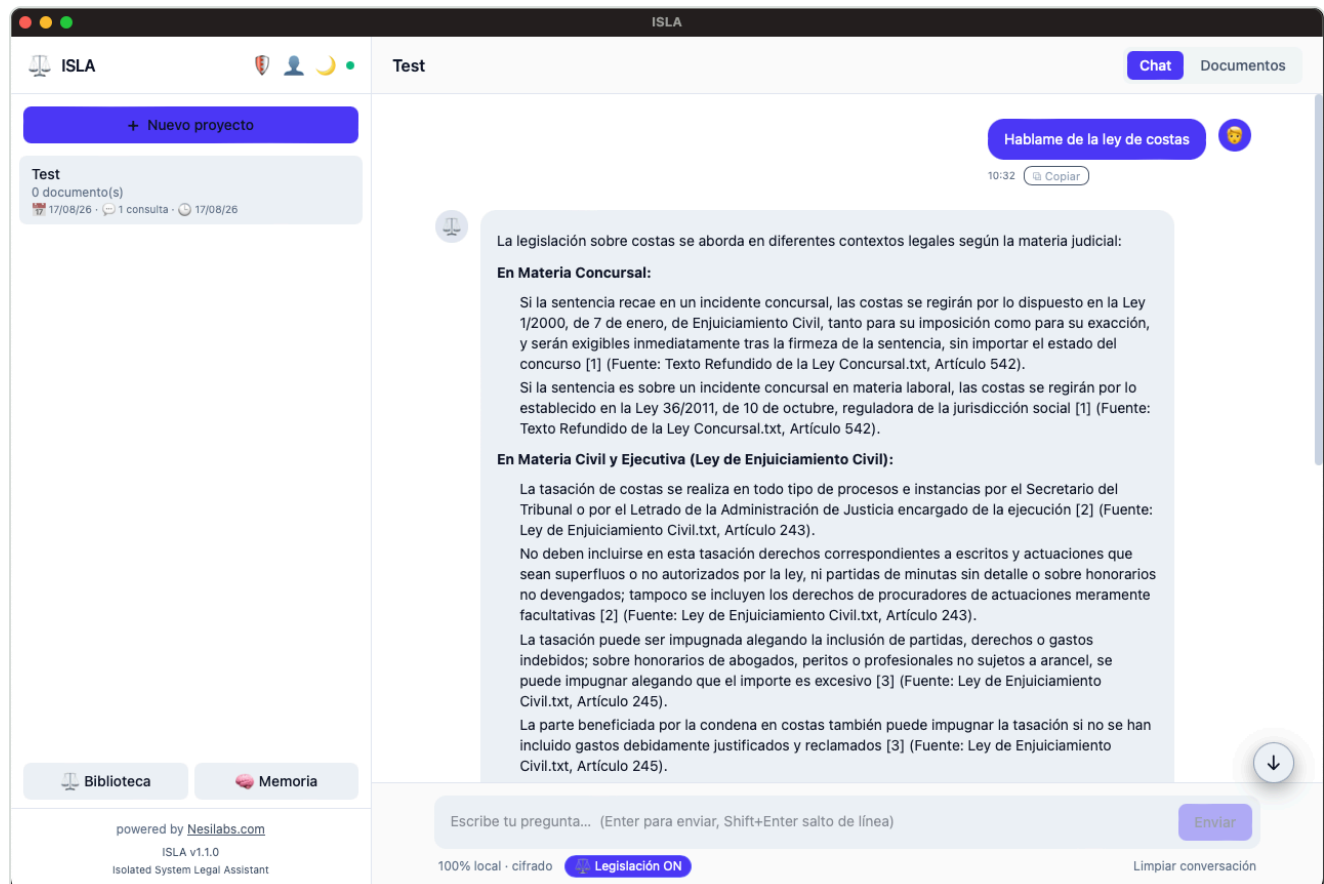


Figure 5. Chat view. An answer about court costs grounded in legislation, citing statute and article (e.g. Civil Procedure Act, arts. 243/245) with per-source [n] markers. Bottom bar: 100 % local · encrypted and the Legislation ON toggle.

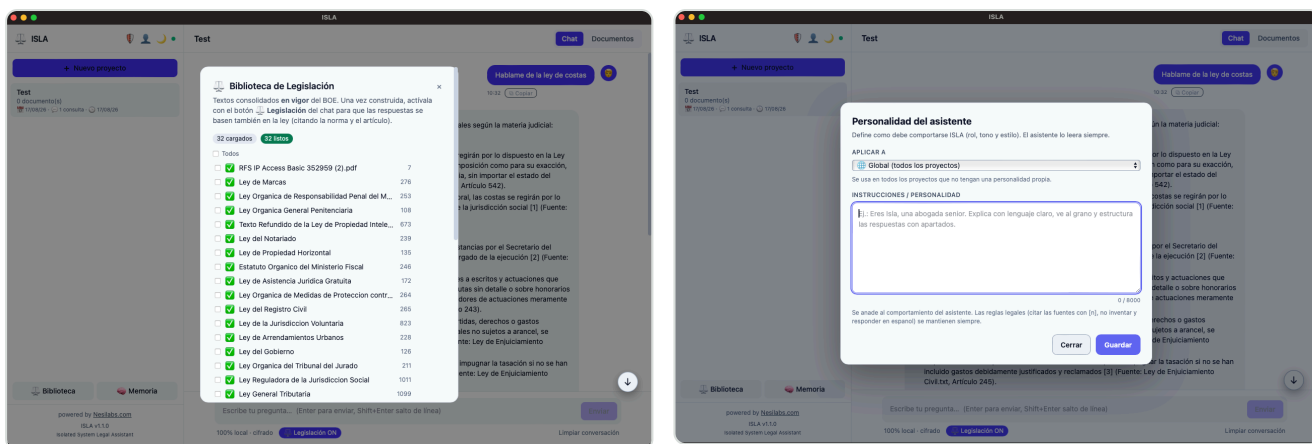


Figure 6. Left: the pre-indexed legislation library (consolidated texts in force from the BOE, with per-statute article counts). Right: the assistant's persona — role/tone instructions applied globally or per project, while the rules to cite sources and answer in Spanish always hold.