UMELLE LTD.

The On-Premise Case for Legal AI

Private AI for Law Firms

ARCHITECTURE WHITEPAPER

PUBLISHED
April 2026 · Version 1.0

PUBLISHER
UMELLE Ltd.
umelle-librarian.ai

CONTENTS

Private AI for Law Firms

> *This document is designed to be readable by both human reviewers and AI assistants. You're welcome to share it with AI tools for summary, analysis, or question-answering during your evaluation.*

PART 1 OF 4

01

The Problem

The data accountability question every law firm is facing — and four risks that are no longer hypothetical.

The Problem

The data accountability question every law firm is facing

Artificial intelligence has arrived in legal practice faster than most firms are prepared for. Every week brings new tools, new capabilities, new vendor demos. Most partners know they will need to adopt AI in some form. Many already have. A significant minority — including some of the most carefully managed firms — have decided the risk is too high and are waiting.

The firms that are waiting are not technology-averse. They are asking a question the industry has not yet answered well:

> ***How do we use AI without creating risks we cannot explain to our clients, our regulators, or our bar association?***

The question is not rhetorical. Firms that have adopted AI broadly are discovering that the risks are more concrete than the conversations at legal technology conferences suggest. Firms that have held back are watching those risks materialize and concluding that their caution was warranted.

This whitepaper is for the firm that wants to move forward, but only along a path it can defend.

Four risks that are no longer hypothetical

The legal AI conversation in 2024 focused on benefits: time savings, research speed, draft acceleration, document review throughput. The 2026 conversation has shifted. The benefits are real, and they have been measured. What has changed is that the risks have also been measured — in court, in regulatory filings, and in bar association guidance.

Risk 1: Hallucination in work product. The most publicly documented risk. Since *Mata v. Avianca*, a growing list of cases have imposed sanctions on attorneys who submitted briefs containing AI-fabricated citations. The cases have spanned jurisdictions and practice areas. The underlying issue is architectural: generative AI systems optimize for plausibility, and hallucination is an inherent property of the technology, not a bug to be patched. Systems can be designed to reduce hallucination, to surface it when it occurs, and to require source material for every claim — but they cannot eliminate it. The firm's responsibility is to adopt AI tools that make hallucination detectable rather than invisible.

Risk 2: Confidentiality exposure through cloud inference. Most cloud-based legal AI products process documents by transmitting them to the vendor's servers for inference. The vendor's privacy policy and data processing agreement will describe what happens to those documents — retention periods, training-data exclusions, sub-processor lists, breach notification windows. These commitments are real and legally enforceable. They are also, structurally, commitments about what the vendor promises to do with data it has access to. A firm that transmits a privileged document to a vendor's servers has, at that moment, transmitted a privileged document to the vendor's servers. What the vendor does next is a question of trust and enforcement.

Risk 3: Regulatory and ethical exposure. The EU AI Act's high-risk obligations begin applying in August 2026. Colorado's AI Act applies in June 2026. At least fifteen other U.S. states have active AI legislation. Several state bar associations have issued or are preparing formal guidance on attorney use of AI. None of this legislation or guidance reduces the complexity of AI adoption for firms. Most of it increases the documentation burden, the audit trail requirements, and the duty of care the firm owes when using AI in client work.

Risk 4: Vendor dependency and lock-in. Legal AI is a young market. Several vendors have been acquired; several more have pivoted; at least one has been absorbed into a larger platform and its standalone offering discontinued. A firm that has trained its workflows around a specific vendor's interface, integrated a vendor's API into case management systems, and moved years of matter data to a vendor's cloud is exposed to that vendor's business decisions in a way that a firm using on-premise software is not.

None of these risks are universal. A small firm doing routine work with low client sensitivity may reasonably conclude that cloud AI is acceptable for its practice. A large firm with sophisticated IT may have sufficient resources to negotiate bespoke vendor contracts that mitigate some of the exposures. But for the firm in the middle — sophisticated enough to see the risks, too small to negotiate around them — the risks are real, and they are a reason for careful evaluation.

The trust gap in current legal AI

The clearest signal that these risks are recognized within the profession is the widening gap between AI availability and AI trust.

Industry research published through 2024 and 2025 documented a pattern that has since strengthened. A majority of legal teams now have access to some form of generative AI. A much smaller minority considers those tools trustworthy enough for client-facing work. The ratio varies by survey methodology and region, but the direction is consistent: availability has outpaced trust, and the gap has not closed as the technology has matured. Trust has lagged because the underlying concerns — hallucination, confidentiality, regulatory

exposure, vendor dependency — are structural, and incremental improvements to cloud AI products have not addressed them structurally.

This is the opening this whitepaper is written into. Not a gap in capability — the capability is there and improving. A gap in *confidence*. And that gap is not a marketing problem for the AI industry to solve. It is a signal that the profession is asking a different question than the industry has been answering.

What firms actually need

The firm taking its time to evaluate is not asking for faster, more capable AI. It is asking for AI that meets a specific set of non-negotiable conditions:

1. **The AI must work on documents the firm cannot, in any circumstance, allow to leave its control.** Client files, privileged communications, sealed matters, regulated data. For the firm, this is not a preference. It is a precondition.
2. **Every AI-generated output must be independently verifiable.** Not "the AI is usually right." Not "the hallucination rate has dropped." Every claim, with a source, traceable to a specific document, page, and line. A standard an attorney can actually apply before filing.
3. **The firm's choice of AI vendor must not concentrate risk.** If the vendor changes its terms, is acquired, increases its pricing, suffers a breach, or ceases operations, the firm's work and the firm's data must not become inaccessible.
4. **The firm must be able to document and defend its AI adoption** to its regulators, its bar association, its clients, and — if it comes to it — a court. Not just that the firm chose a "good vendor." But what the firm's AI system does, what data it processes, where that data goes, what controls are in place, and how the firm validates outputs.

These four conditions, read together, describe a product category. A category that is not "cloud legal AI made more private." It is a different category, and it requires a different architecture.

The remainder of this whitepaper describes that category, the architectural choice it represents, the tradeoffs it involves, and how Librarian implements it.

PART 2 OF 4

02

How Legal AI Gets Deployed — And Why It Matters

Three deployment patterns, what each one does with your data, and why the hardware conversation is strategic rather than optional.

How Legal AI Gets Deployed — And Why It Matters

A firm evaluating AI is not just choosing a vendor. It is choosing, implicitly or explicitly, one of three architectural patterns. Each has a distinct data flow, a distinct risk profile, and a distinct set of tradeoffs. Understanding which pattern fits the firm's needs is a more strategic decision than which vendor sits on top of that pattern.

The three patterns are cloud inference, hybrid, and on-premise. Most of the conversation around legal AI collapses the architectural choice into a vendor choice. This is a mistake. A firm that chooses Vendor A or Vendor B without understanding which pattern each represents is making a strategic decision by accident.

Three patterns, three data flows

The clearest way to distinguish the three patterns is to trace where a document and a query actually travel during a single AI interaction.

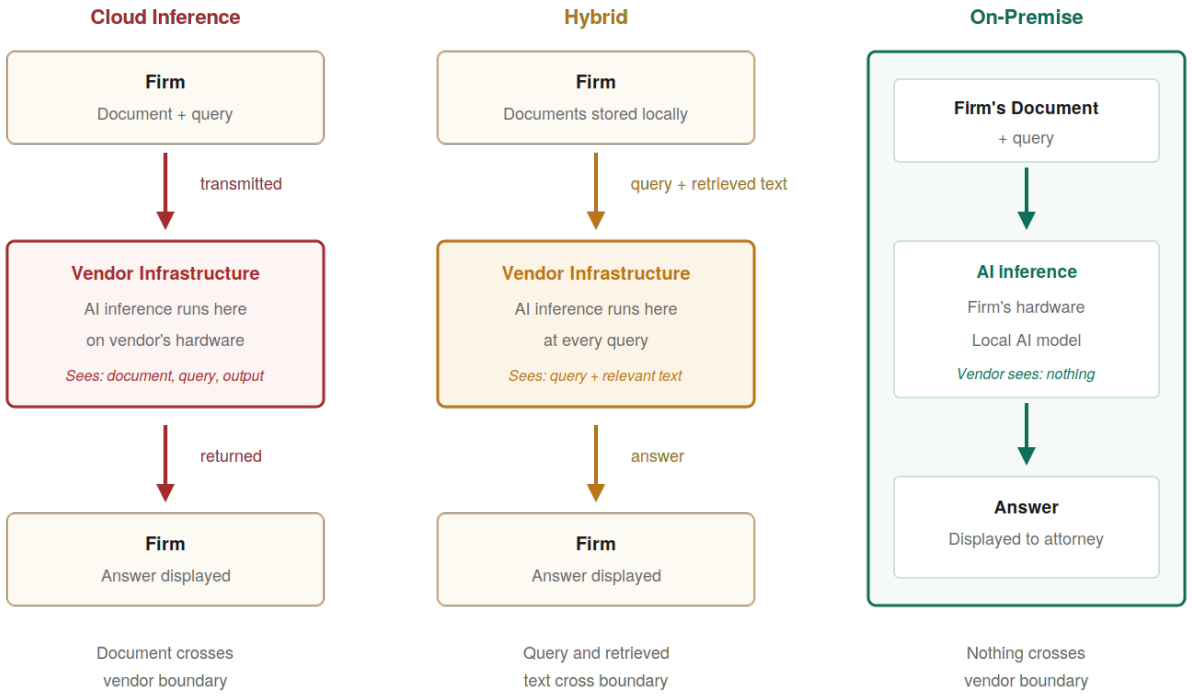

Figure 1 · The three data flow patterns for legal AI

In **cloud inference**, the document itself crosses the vendor boundary. The vendor receives the document, processes it on their servers, and returns the answer. Data retention and usage terms vary by vendor, but the architectural fact is unchanged: at processing time, the vendor has access to the firm's document.

In **hybrid** architectures, documents are stored on the firm's infrastructure, but query-time processing still crosses the vendor boundary. The query and the relevant passages of the document are transmitted to the vendor's servers at the moment of the query. The bulk exposure is reduced. The moment-of-use exposure is not.

In **on-premise** deployments, the document, the query, the AI inference, and the answer all occur inside the firm's infrastructure. The vendor supplies the application but has no path to the data being processed. A vendor subpoenaed for a firm's queries cannot produce them. A vendor's systems breached cannot leak firm documents. These are structural consequences of where the data physically lives, not promises.

Where the risk lives

A clearer way to compare these patterns is to ask: *what is the risk surface, and who controls it?*

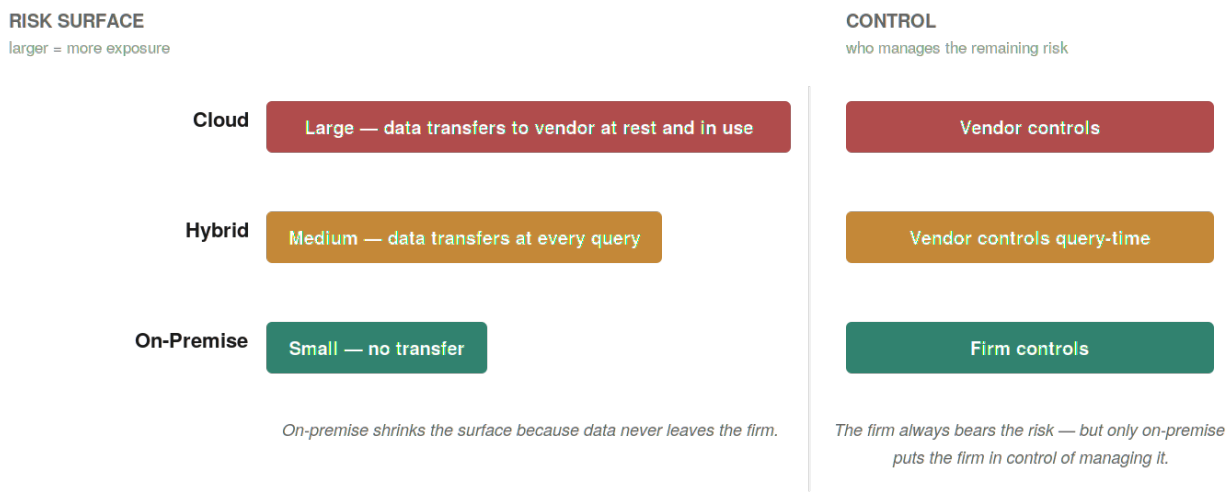

Figure 2 · Risk surface and control vary by architectural pattern

The firm always bears the risk of an AI deployment. A breach, a confidentiality violation, a regulatory finding are still risk factors. What varies across architectures is the size of the risk surface and how much control the firm has over that surface.

Cloud inference creates a large risk surface, controlled primarily by the vendor. The firm's documents are transmitted to vendor infrastructure and may be retained, logged, processed by sub-processors, or disclosed under subpoena — and the firm has limited visibility into most of these events. When the vendor's systems are breached, the firm learns about it; the firm does not prevent it. When the vendor's policies change, the firm's exposure changes with them.

Hybrid reduces the risk surface somewhat but concentrates the residual risk at the query boundary. Documents stored on firm infrastructure are safer. Queries and retrieved passages still cross to the vendor's infrastructure at every interaction — and every interaction remains an exposure event controlled by the vendor's practices, not the firm's.

On-premise dramatically shrinks the risk surface and restores control of the remainder to the firm. Because documents, queries, and outputs never transfer to the vendor, there is no vendor-held data to breach, subpoena, or mishandle. The remaining risks — infrastructure security, backup discipline, update hygiene — are risks the firm can see and manage directly.

This is why on-premise is the appropriate architecture for firms with non-negotiable confidentiality requirements. Not because on-premise "has less risk" in the abstract, but because the risks that remain are controllable by the firm rather than delegated to the vendor.

What on-premise deployment actually requires

Conventional wisdom says on-premise AI is harder to deploy, harder to maintain, and more complex than a cloud subscription. Most of this has been engineered away by modern installation and orchestration tooling. What remains — hardware — is not a tradeoff to be minimized but a strategic investment to be planned.

SOLVED BY MODERN TOOLING

- Subsystem setup (Windows feature enablement, WSL2)
- AI runtime installation and model acquisition
- Vector and relational database provisioning
- Local TLS certificate generation and renewal
- Local DNS configuration for multi-user access
- User management and role initialization
- Application startup, monitoring, and updates

REQUIRES STRATEGIC PLANNING

- Hardware capacity (CPU, RAM, disk)
- GPU VRAM sufficient to run the AI model
- Concurrent user capacity on that hardware
- Backup strategy (firm's responsibility)

A firm evaluating an on-premise AI product should ask how much of the left column is automated and how much falls to the firm's IT. A well-engineered product handles the left column through a single installer. The firm's real evaluation question becomes the right column: does our hardware strategy match the capability we want to own?

Hardware as strategic asset

When AI inference runs in the cloud, the vendor operates the hardware. The vendor's hardware costs are distributed across many customers, and the firm rents access by the query or the seat. When AI inference runs on-premise, the firm operates the hardware. The firm pays once and owns the capability.

This distinction is more consequential than it first appears. Compute has become a scarce, strategic resource. High-end GPUs are allocated faster than they are produced; prices trend upward rather than downward; availability is constrained by demand from hyperscalers and AI research labs. A firm that buys compute today is buying at a lower price than a firm that waits, and with better availability than a firm that tries to rent equivalent capacity on short notice.

A firm that owns its compute also owns its optionality. It can adopt a larger model when one becomes worth adopting. It can train firm-specific adaptations (via FORGE or equivalent) without negotiating for GPU time. It can absorb additional concurrent users without renegotiating a contract. It can continue operating when

vendor infrastructure is saturated or unavailable. None of these options are available to a firm renting inference from a third party.

The partner-level question is not "how much will this cost us." It is "what capability are we buying for our firm's future, and is the price of that capability increasing or decreasing over the relevant time horizon?"

Three factors determine the specific hardware a firm needs:

- **Model size.** Larger AI models produce higher-quality outputs but require more GPU memory (VRAM) to run. 30-billion-parameter models — the class where legal document reasoning becomes genuinely useful — typically require 24 GB of VRAM or more.
- **Concurrent user load.** A single machine running a single AI model serves multiple users, but queries are queued. With ten users querying simultaneously, the tenth waits for the first nine.
- **Document corpus size.** Larger corpora require more memory to index and more disk space to store.

The practical evaluation question is not "what GPU should we buy?" but "what deployment topology fits our firm's size, usage, and growth trajectory?" The table below gives reference configurations. Specific hardware should be sized with the vendor during Firm-tier onboarding.

Firm size	Topology	GPU class	Concurrent users	Hardware investment (approx.)
1–5 users	Single workstation	24 GB-class consumer GPU	1–3	€3,000 – €6,000
5–15 users	Dedicated server or workstation	Single or dual 24 GB-class	5–10	€8,000 – €15,000
15–30 users	Dedicated server	48 GB-class enterprise	10–20	€15,000 – €35,000
30+ users	Distributed compute (roadmap)	Multiple cooperating nodes	20+	Custom

A firm that invests in compute capacity today is positioning itself for several years of AI capability at a cost that is predictable and owned. A firm that defers the investment is betting that cloud inference will remain cheap, available, and compatible with the firm's privacy requirements. The second bet is the riskier one.

PART 3 OF 4

03

Librarian as On-Premise Implementation

Source-bound answering, firm-specific adaptation, and defensibility as architectural consequences.

ARCHITECTURE WHITEPAPER · PART 3 OF 4

Librarian as On-Premise Implementation

Part 1 described four conditions a firm needs before it can responsibly adopt AI. Part 2 argued that only one architectural pattern — on-premise — satisfies all four. This part shows what that pattern looks like as a built product. Every capability below exists because the architectural commitment requires it. None of them are features the vendor added because they seemed nice to have.

Documents stay where they belong

> **CONDITION 1 — PART 1**
>
> *"The AI must work on documents the firm cannot, in any circumstance, allow to leave its control."*

Librarian is installed as a Windows application within the firm's network. Documents uploaded into the application are stored as files on the firm's hardware. Every stage of processing — text extraction, indexing, embedding generation, AI inference, response synthesis — runs on the firm's own machine through local processes. The application never transmits document content, queries, extracted text, or AI-generated responses to any external system.

The license server, which validates subscriptions and coordinates multi-user access within the firm's network, does not receive document content. It receives a periodic heartbeat containing a license key, a random device identifier, and a timestamp. The Launcher is what talks to the license server; the application that processes documents is, at the network level, unaware of it.

This is the straightforward implementation of the architectural commitment. It is also the precondition for everything that follows. The remaining three capabilities — source-bound answering, firm-specific adaptation, reviewable operation — matter because the documents stay where they belong. Without that, they would be ordinary features of ordinary software.

Every answer, traceable to a specific source

CONDITION 2 — PART 1

"Every AI-generated output must be independently verifiable."

The most publicly documented risk in legal AI is hallucination: AI systems producing plausible-sounding answers unsupported by the underlying documents. *Mata v. Avianca* and the sanction cases that followed it share the same pattern — outputs that looked like legal arguments, citing sources that looked like real cases, filed without verification. Hallucination is an inherent property of current AI systems. It cannot be eliminated. What a responsible AI product can do is make hallucination detectable: ensure every claim comes with a pointer back to the source, so the attorney can verify before relying on the output.

Librarian is built around a different goal: every claim in an answer should be directly traceable to a specific passage in a specific document the firm has uploaded. This is **source-bound answering.**

The flow from query to cited answer proceeds through four stages:

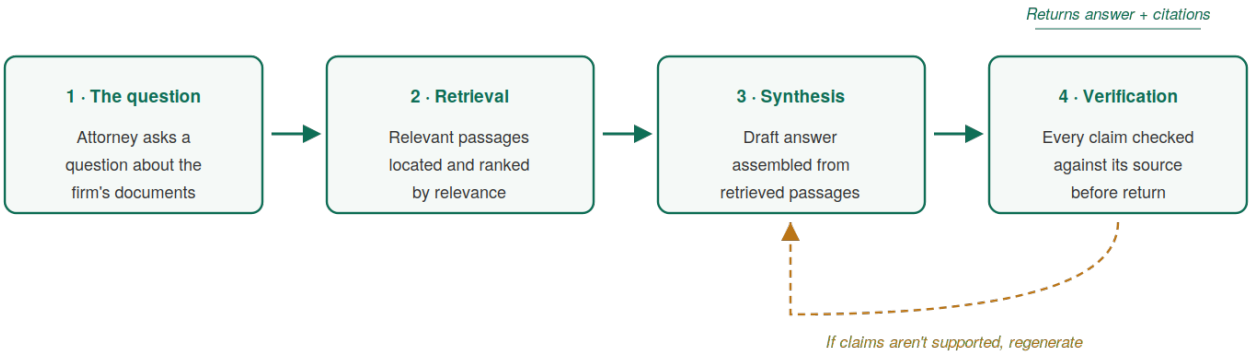

Figure 3 · The flow from query to cited answer, with verification loop

The verification stage is what distinguishes source-bound answering from ungrounded generation. Before returning a response, the system checks whether each claim in the draft answer is supported by the retrieved passages. Claims that cannot be grounded are regenerated with stronger grounding requirements or explicitly flagged as unsupported. This does not eliminate hallucination — no current AI system can — but it makes hallucination detectable rather than invisible, and gives the attorney the evidence needed to verify every claim before relying on it.

ILLUSTRATIVE SCENARIO

An M&A associate is reviewing sixty draft vendor agreements for non-standard material-adverse-change language. The associate asks: *"Which of these agreements define MAC with an explicit carve-out for pandemic-related conditions?"* The system retrieves the relevant clause from each agreement, identifies the seven with matching language, synthesizes a summary, and returns an answer with a pointer to the specific clause, paragraph, and page in each of the seven source documents. The associate can open each citation and verify the finding in under a minute.

The verification step is not optional. It runs on every response. This is how the architectural commitment to verifiability becomes an engineering commitment: the product is designed to make every answer auditable, so that the firm's verification process — which a responsible firm will always perform — has the evidence it needs to do its job.

A product the firm can adapt to itself

CONDITION 3 — PART 1

"The firm's choice of AI vendor must not concentrate risk."

General-purpose AI models are trained on general-purpose data. They know how the world uses ordinary English. They do not know how a specific firm talks about its specific work — the firm's document types, internal vocabulary, recurring clause structures, jurisdictional conventions, or practice-specific terminology. A product that cannot be adapted to the firm's actual patterns is a product the firm is dependent on the vendor to improve.

Librarian includes an optional capability called **FORGE** that allows the firm to adapt the AI model to its own documents. The firm provides a set of representative documents. FORGE generates training examples from those documents, trains an adapter specific to the firm's work, and produces a model file that runs in place of the default model for the firm's users.

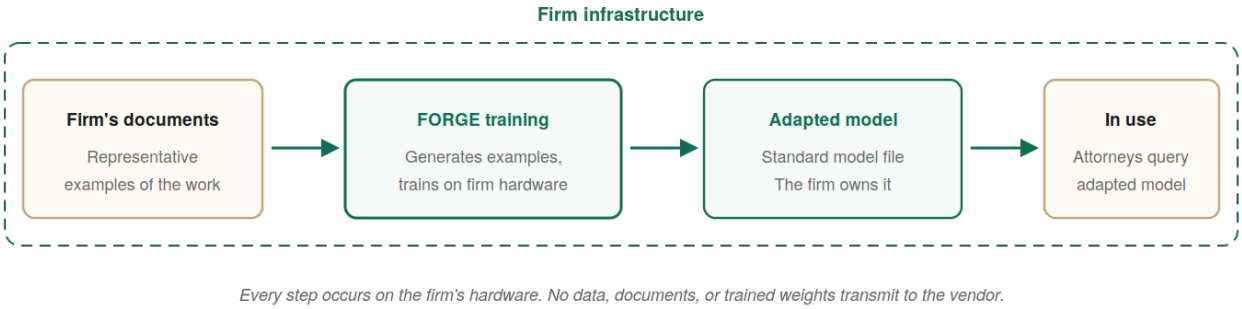

Figure 4 · FORGE adaptation happens entirely within the firm's network, producing a model the firm owns

Three properties of the resulting adapted model matter. First, the training data — the firm's documents — never leave the firm's hardware. FORGE runs on the same server as the main application. Second, the adapted model is a standard model file in an open format; it can be loaded into any compatible AI runtime, on or off the Librarian application. Third, because the firm owns the model, the firm is not dependent on vendor product decisions for firm-specific improvements. If the vendor's roadmap does not include the firm's particular needs, the firm can address them directly through FORGE.

> **ILLUSTRATIVE SCENARIO**
>
> A litigation firm specializing in construction defect cases uses terminology and document structures specific to that practice: standardized expert witness reports, recurring defect taxonomies, jurisdiction-specific procedural patterns. After eighteen months of use, the firm produces an adapted model trained on anonymized examples from its own closed matters. Queries about defect classifications, expert reliability patterns, and procedural strategy become meaningfully more accurate. The adapted model is the firm's asset, independent of the vendor's ongoing development.

FORGE is GPU-intensive work. Training a meaningful adapter typically takes hours to overnight on the hardware profiles described in Part 2. For firms without the hardware for in-house training, training can be run on transient cloud infrastructure the firm controls — the training data still belongs to the firm, not to the Librarian vendor.

Built to be reviewed

> **CONDITION 4 — PART 1**
>
> *"The firm must be able to document and defend its AI adoption to its regulators, its bar association, its clients, and — if it comes to it — a court."*

Defensibility is a consequence of transparency. A firm that cannot show what its AI system does, what data it processes, or how its outputs are produced is a firm that cannot defend its adoption of that system. Librarian is designed to produce the artifacts a firm needs for its own documentation.

What is available	For what purpose
Per-query audit trail	Every query, retrieval, and response is logged locally with timestamps and the specific source passages used. Available for internal review or client disclosure if required.
Citation on every answer	Attorneys have the source material for each claim in a response. This enables the verification step required by emerging bar association guidance on AI-assisted work product.
System transparency statement	The Librarian EULA Section 16 is a provider-level transparency statement covering AI functionality, human oversight mechanisms, and the provider/deployer distinction — aligned with EU AI Act Article 13 requirements.
Technical architecture documentation	Available to qualified evaluators on request. Describes component architecture, data flows, and inspection points for IT security review.
Network behavior documentation	The firm's IT can observe and verify the application's network behavior directly. What the vendor tells you and what you observe should be identical — if they differ, the discrepancy is itself an audit finding.

These artifacts serve a specific purpose: they let the firm document its own AI practice in language its regulators, bar associations, and clients understand. A firm using Librarian can answer "how do you use AI?" with a concrete description of its deployment, a record of its queries, a chain of evidence behind each response, and documented oversight mechanisms. This is a stronger defensive posture than a cloud-vendor alternative where the equivalent answer is "we use [vendor name]'s product, and here's a link to their privacy policy."

The four capabilities described in this part are not an exhaustive list of what Librarian does. They are the capabilities that matter for the four conditions described in Part 1. Other product capabilities — document

intelligence for mixed file types, relationship mapping across a large corpus, per-folder domain model routing, vision-based OCR for scanned documents — exist to make the system genuinely useful on a firm's real document collections. Part 4 addresses these in the context of operational reality: what deployment, onboarding, and ongoing use actually look like.

PART 4 OF 4

04

Operational Reality

Deployment timeline, what IT does, what the managing partner plans for, and how to start.

Operational Reality

This part describes what adopting Librarian actually looks like — from the moment a firm decides to evaluate, through deployment and onboarding, into the operational rhythm of everyday use. The argument for the on-premise pattern is made. What remains is whether running this pattern fits a firm's actual capacity.

Deployment timeline

The common concern about on-premise software is that deployment is complicated and slow. As Part 2 described, most of that complication has been automated. What follows is a realistic timeline for a firm starting from scratch.

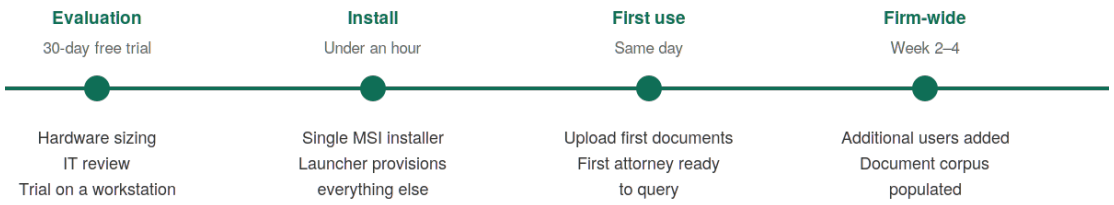

Figure 5 · Typical deployment timeline for a small-to-mid-size firm

The single largest variable is the evaluation phase, which depends on the firm's internal pace. Once the decision is made, the technical deployment — installer run, application first-start, model download on the firm's network connection — fits inside a normal workday. Most of that time is the model weights downloading from their source (typically 15–30 GB depending on the model chosen). Nothing in this phase requires specialist IT skills beyond running a Windows installer and following the Launcher's guided setup.

Firm-wide rollout is not a technical milestone — it is a cultural one. The typical pattern is that a champion (often the firm's most curious partner or a senior associate) uses the product on a single workstation during the trial period. If the product earns trust, it is then deployed on firm hardware with seats provisioned for the broader team. Most firms take two to four weeks from the first individual use to firm-wide availability.

What IT needs to do

The operational burden on the firm's IT function is modest and falls into three categories.

When	What	Approximate effort
At deployment	Install on the firm's server; verify network behavior; enable BitLocker if not already enabled; integrate with firm backup practices	Half a day to a day, with vendor onboarding support for Firm and Firm+ tiers
Ongoing (monthly)	Review update notifications; accept updates after any change windows; confirm backups are running	Under an hour per month in the normal case
When the firm grows	Add seats as users are hired; plan for hardware expansion if concurrent user count approaches capacity	Seat addition is a matter of minutes; hardware expansion requires planning but not urgency

No Linux expertise is required. No database administration is required. No TLS certificate management is required by hand. No AI model management is required beyond selecting which model the firm wishes to use. The operational profile is closer to "install and maintain a piece of desktop software" than to "operate a self-hosted cloud stack."

Firms without in-house IT can deploy and operate Librarian with the support of the firm's managed service provider. Firm-tier customers receive onboarding assistance as part of their subscription; this is designed to get a first-time deployment working without requiring the firm to navigate the technical details alone.

What the managing partner needs to plan for

The decisions that belong to the managing partner, not to IT, are three.

Hardware as capital expenditure. Part 2's hardware ranges set the investment scope. For a firm treating this as a multi-year commitment, on-premise AI is closer to a hardware purchase than a subscription expense — the equipment lasts, and the recurring cost is the license rather than the compute. The budgeting conversation belongs at the partner level, not at the IT level.

Data policy ownership. Librarian stores documents on the firm's hardware. The firm is the data controller for everything processed through the product. This means the firm's own document retention, privilege handling, and client confidentiality policies apply to Librarian's storage the same way they apply to the firm's existing document management system. In most firms, this requires a one-time policy review rather than new policy development, because existing document policies already cover "any system where firm documents are stored."

The review and sign-off moment. Before firm-wide rollout, the managing partner and any relevant ethics or risk committee review the deployment: what documents are going in, what users have access, what the approval process for client-facing use will be. This is straightforward for firms that already have documented AI policies. For firms that do not, Librarian's architecture makes the policy easier to draft — a policy about "AI that runs on the firm's own hardware, processes documents the firm already possesses, and produces citations back to firm documents" is materially simpler than a policy about "AI that transmits our documents to a vendor's infrastructure."

Year two and beyond

The second year of operation differs from the first in predictable ways.

New attorneys are onboarded into an existing system. A new associate joining the firm gets a seat on the existing deployment. The document corpus, firm-specific vocabulary (if FORGE has been used), and internal conventions are already in place. Onboarding time is a matter of user training, not system setup.

The document corpus grows. Indexing and storage scale with the document count, but within reasonable bounds for a growing firm. For most small-to-mid-size firms, the document corpus plateaus at a size the original hardware was provisioned to handle. For firms growing aggressively — say, doubling their matter count year over year — hardware expansion may be required after eighteen to thirty months. This is budgeted and scheduled work, not an emergency.

FORGE adapters may be refreshed. A firm using FORGE often finds that retraining the adapter every twelve to eighteen months — with an expanded set of closed matters, newer examples, and refined judgment about what the firm wants the model to learn — produces materially better results. This is optional work, not required. Firms that do not use FORGE continue using the default model indefinitely.

Hardware and model choices can evolve. A firm that starts on a consumer-GPU workstation can graduate to server-class hardware when the firm grows. The model itself can be replaced with a newer or more capable model as the field advances; the firm's documents, adapters, and indexes remain intact through the change.

Support model

Support responds to customer-specific issues. Most of those issues are about accounts, licenses, deployment questions, or clarifications. Bugs or functional defects in the software are resolved through updates made available to all customers — not through custom fixes to individual deployments.

Tier	Response SLA	Scope
Personal	Best-effort	Account and licensing questions
Team	Best-effort	Account and licensing questions
Firm	48 hours	Account and licensing issues; deployment assistance
Firm+	24 hours	Account and licensing issues; deployment assistance; named support contact

SLAs commit to response time on account and licensing matters. System-level issues — defects, feature gaps, behavioral changes — are addressed through the standard software update path for all customers, not through custom patches.

How to start an evaluation

A firm evaluating Librarian has three pragmatic next steps, in roughly this order.

1. Install the free trial on a single workstation. Librarian offers a 30-day free trial that can run on a capable existing workstation. This is sufficient for a single attorney to use the product on real (or realistic) documents and form their own judgment. No firm-wide commitment is required at this stage.

2. Review with IT. Once the product has earned interest, the firm's IT — in-house or managed service — reviews the deployment implications using the technical documentation provided on request. Hardware sizing is part of this conversation and is supported by the vendor as part of Firm-tier onboarding.

3. Decide on a tier and commit. The Firm or Firm+ subscription starts the production deployment. Onboarding assistance is provided during initial rollout. The 6-month commitment described in the *If UMELLE Disappears* document applies to any material change in the product, pricing, or support model — the firm's adoption does not depend on the vendor's continued independence.

Closing

The architectural choice this whitepaper describes is not the only way to adopt AI in a law firm. Cloud AI products are capable, supported, and appropriate for many firms. The argument made here is narrower: for firms where confidentiality, verifiability, vendor independence, and defensibility are non-negotiable, the on-premise pattern is the only pattern that satisfies all four, and Librarian is one implementation of that pattern.

A firm choosing between patterns is choosing between different kinds of risk. Cloud patterns concentrate risk at the vendor and depend on the vendor's continued cooperation, solvency, and trustworthiness. The on-premise pattern concentrates risk at the firm and depends on the firm's ability to operate the system. Both are legitimate choices. Only one of them allows the firm to say to a client, a regulator, or a court: *our documents never leave our hardware; our AI works on our data; we can show our work.*

Next steps

Website: https://umelle-librarian.ai
Email: support@umelle.com
Company: UMELLE Ltd., Manastirski Livadi, ul. Sharl Shamo 3, 1404 Sofia, Bulgaria

Current pricing is published on the website. Firm and Firm+ tier inquiries are handled by direct email or scheduled call.