

Private Legal AI

Solution & Architecture Brief

AI deployed inside your infrastructure — with matter-aware retrieval, cited answers, and no client documents sent to external AI providers.

Firm-controlled

Model-agnostic

Citations

Matter isolation

For law firms that want AI capability without routing confidential work through a third-party inference service.

The problem isn't AI capability. It's where the data goes.

General-purpose cloud AI is useful. But legal work adds constraints that ordinary SaaS assumptions do not solve.

- Client documents may contain privileged, regulated, or highly sensitive information.
- Matter access is often narrower than firm-wide document access.
- Firms need defensible controls around retention, logging, identity, and updates.
- Security teams need to know exactly where prompts, embeddings, and source text live.

LEXDEPLOY APPROACH

Bring the AI to the data.

The inference, retrieval, and indexing stack runs on firm-controlled infrastructure. Client documents do not need to leave that environment for routine AI use.

No per-query token charges

Local data plane

On-premises deployment strengthens data-sovereignty and confidentiality controls; it does not, by itself, determine legal privilege status.

Useful AI starts with work lawyers already do.

LexDeploy is designed around source-grounded legal workflows, not generic chat.

01 Search with citations

Ask natural-language questions across a matter and return answer passages with document/page provenance.

02 Compare drafts

Identify changed clauses, defined terms, obligations, and deviations across document versions.

03 Matter summaries

Create structured timelines, issue lists, and factual summaries grounded in the selected matter.

04 Deposition & hearing prep

Surface relevant testimony, exhibits, and prior statements for attorney review.

05 Discovery review

Find recurring topics, entities, dates, obligations, and potentially relevant document clusters.

06 Internal knowledge

Search approved precedent, playbooks, policies, and templates without commingling restricted matters.

Outputs support legal professionals and require appropriate human review; the system is not a substitute for professional judgment.

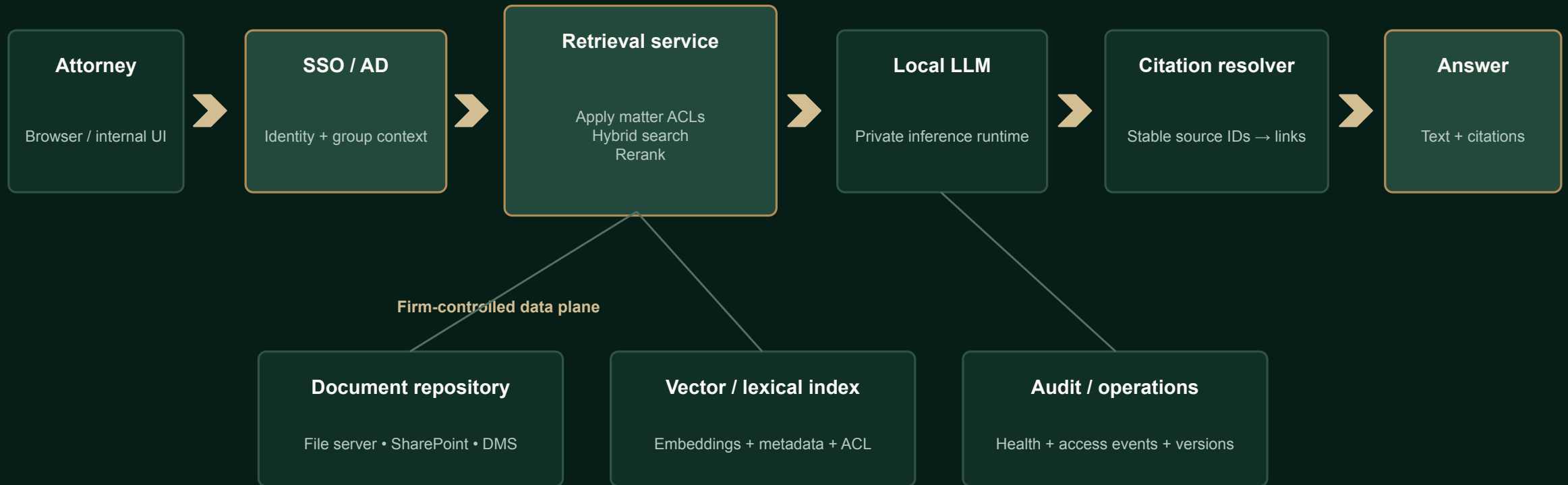
How the system turns firm documents into cited answers.

The retrieval layer narrows the model’s context to relevant, authorized source material.



Authorization filtering occurs before semantic retrieval so inaccessible content is not retrieved and then “hidden” after the fact.

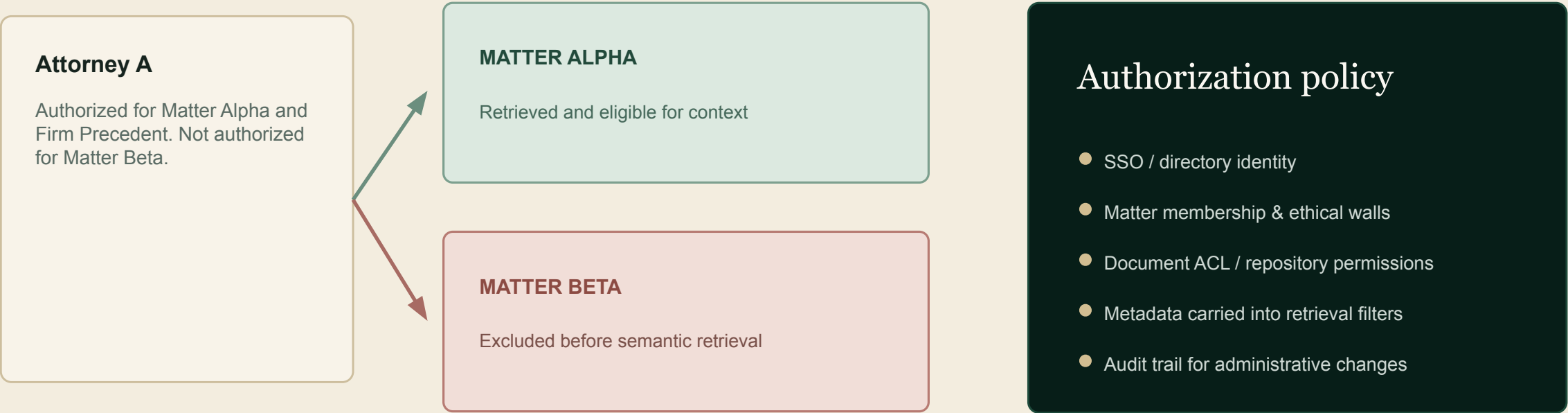
Query path: identity, authorization, retrieval, inference, citation.



NO CLIENT CONTENT → EXTERNAL AI CLOUD

Matter isolation is a retrieval control, not a disclaimer.

The user’s authorization context constrains search before any source text reaches the model.



Embeddings and vector metadata are treated as sensitive derived data and protected with the same access-control discipline as source documents.

Data sovereignty by design.

Confidential content stays in the firm-controlled data plane; operational management is deliberately separated.

1 Client content

Documents, prompts, retrieved passages, embeddings, and answers remain within firm-controlled infrastructure.

2 Identity & encryption

SSO/AD, TLS on the internal network, encryption at rest, encrypted backups, and managed secrets.

3 Operational telemetry

GPU/CPU/RAM/disk, service health, queue/latency, index health, and backup status — not prompts or document text.

4 Controlled change

Signed/versioned releases, model hashes, tested upgrade packages, rollback points, and offline update options.

Security architecture is part of the product — not an implementation afterthought.

Two operating modes for different security requirements.

Managed Sovereign

Recommended default

- Firm-owned / firm-controlled infrastructure
- Restricted, audited remote administration
- Operational telemetry only — no client content
- Tested model/software updates and rollback
- Break-glass access only with firm approval

Best for most firms

Fully Isolated

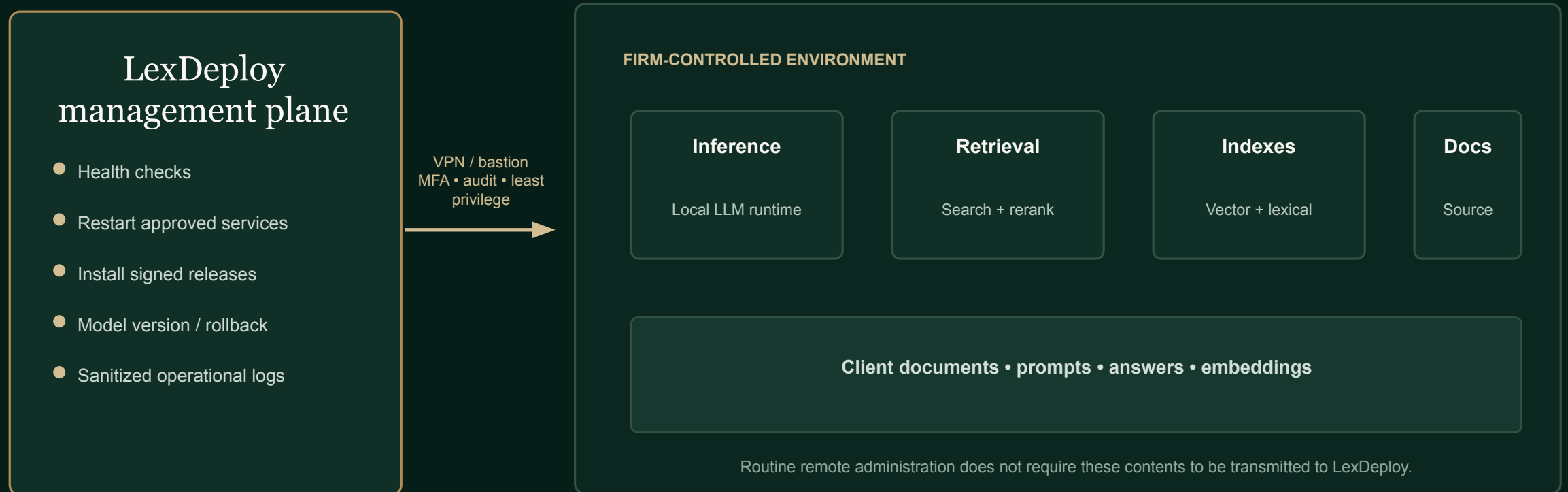
Highest isolation

- No persistent LexDeploy remote access
- No external runtime dependency
- Signed offline release packages
- Updates applied by firm IT or scheduled maintenance
- Operational burden shifts toward customer IT

For highest-security matters

Remote operations without routine content access.

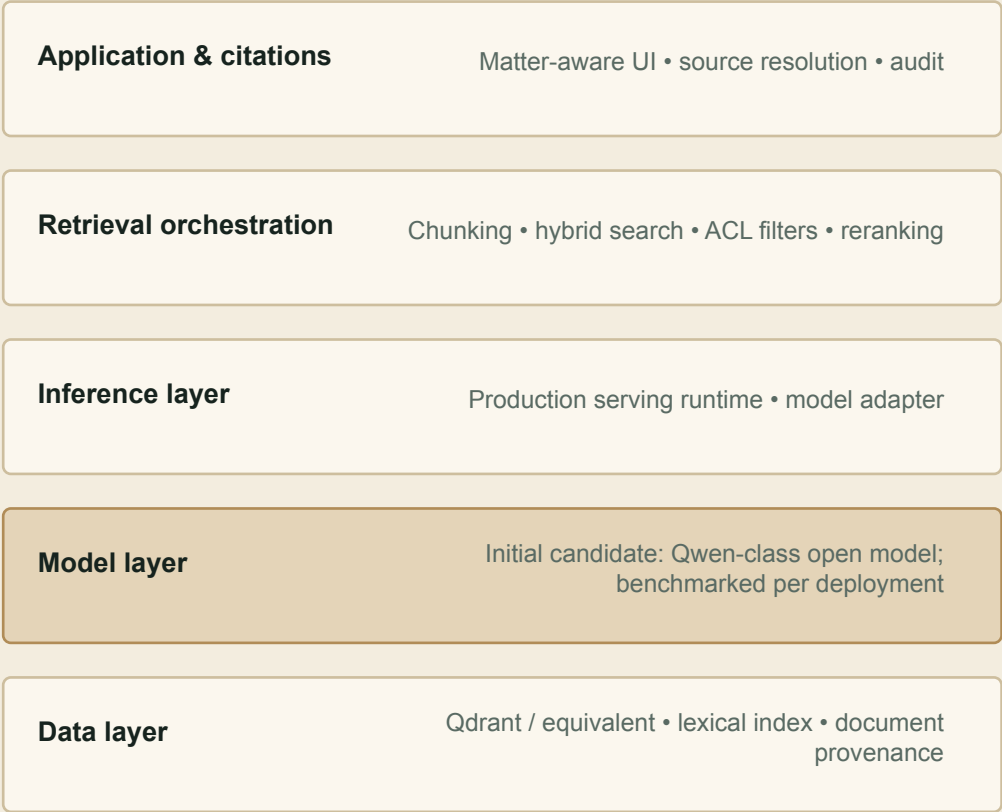
A separate management plane handles service health and approved change; the AI data plane stays inside the firm.



Permanent unrestricted root access is avoided; severe incidents use separately governed break-glass access with firm approval and full audit logging.

The product is the private platform — not one model.

Model, embedding, reranking, and serving components can evolve independently as quality and hardware change.



Why this matters

- Hardware fit**
The best model depends on GPU memory, concurrency, context, and latency targets — not benchmark headlines.
- Upgrade path**
Models can improve without redesigning identity, retrieval, provenance, or document controls.
- Benchmark first**
Candidate models are evaluated on the firm's representative legal tasks before production sizing.

Prove quality and controls before buying production hardware.

The POC is a controlled evaluation, not a demo with cherry-picked prompts.

01	02	03	04
Select corpus	Build question set	Run evaluation	Decide
Representative documents, versions, scans, tables, and permission boundaries.	75–150 representative questions: exact clause, defined term, cross-document, negative, multi-hop, unauthorized matter.	Retrieval quality, answer correctness, citation coverage, unsupported claims, latency, and permission tests.	Use measured results to choose model, context window, GPU sizing, integrations, and production scope.

Acceptance targets include: unauthorized retrieval = 0 • citation correctness • unsupported-claim rate • Recall@K • p50/p95 latency

\$0 proof of concept

Production scope depends on integration and security complexity.

Typical deployment

Lower-complexity baseline

Microsoft Entra ID / Active Directory
Standard file repository or SharePoint
Single production server
Managed Sovereign operations
Standard matter / group access mapping
Routine backup and recovery design

Base implementation range

Complexity drivers

Advanced scope

iManage / NetDocuments integrations
Complex ethical walls and nested ACLs
High availability / redundant infrastructure
Large legacy ingestion / OCR remediation
Air-gapped or fully isolated operating model
Custom workflows, retention, or audit requirements

Quoted after assessment

LexDeploy implementation is priced on integration and security scope rather than as a percentage of hardware cost.

Simple economics. Hardware stays yours.

No hardware markup and no per-query token charges.



Additional capacity, advanced integrations, high availability, and fully isolated deployments are scoped separately.

What the monthly service includes — and what it doesn't.

Included

- Service health and resource monitoring
- Security patches and tested software updates
- Model updates with evaluation / rollback
- Retrieval and index health checks
- Backup verification and performance tuning
- Business-hours operational support

Scoped separately

- New DMS or line-of-business integrations
- Major custom workflow or product features
- Large data migrations / remediation projects
- Hardware replacement or expansion
- After-hours emergency SLA coverage
- Material architecture changes / HA redesign

The service boundary keeps the platform maintainable while making larger changes explicit rather than hiding them inside an ambiguous retainer.

Start with a controlled evaluation.

A no-cost Private AI Assessment defines the representative documents, permission boundaries, questions, and success criteria for a proof of concept.

REQUEST A PRIVATE AI ASSESSMENT

No production commitment. No hardware purchase before the POC. No need to send client documents to an external AI service.

