ALLIED BizTech SOLUTIONS

FIELD GUIDE · 2026

Rebuild your firm's legal AI yourself. *The plan we follow.*

For AmLaw 200 and AmLaw 100 firms paying per seat for legal AI who want to own the workflows, keep every document in their own tenancy, and keep the research sources they trust.

- → Is it worth it, at your size
- → What to rebuild and what to keep
- → The architecture, in three layers
- → Rollout and go-live checklists
- → The matter-context model
- → A week-by-week build plan
- → Privilege and confidentiality
- → Where firms get stuck

Legal AI is priced per seat. *The models, per document.*

Harvey enterprise runs ~$3,300 an attorney a year. The models underneath cost $0.01–0.05 per document analyzed. At 50 attorneys, that is ~$165K a year to wrap APIs your firm can call direct for ~$6K a year.

THE LEGAL STACK	SEAT PRICE	UNDERNEATH	VERDICT
Legal AI · Harvey, Spellbook, Casetext	~$3,300 an attorney a year on Harvey; Spellbook also per seat	Foundation models at $0.01–0.05 a document: 95%+ wrapper margin	**Rebuild first**
Matter and billing · Clio, Filevine	$60–$120 a seat a month	A matter app on Postgres, QuickBooks or Sage for billing: 60–75% recovered at 30+ attorneys	Rebuild next
Documents · NetDocuments, iManage	$80–$120 a seat a month	S3, Postgres and Meilisearch at ~$8–15K a year, at any seat count	Connect now
Research data · Lex Machina, Bloomberg Law	Specialized data feeds	The data itself	Keep

The AmLaw 200 case: 50 attorneys

BEFORE	A YEAR	AFTER THE REBUILD	COST
Harvey AI, 50 attorneys	$142K	Anthropic + OpenAI direct	~$6K a year
		AWS hosting + Postgres	~$3.6K a year
Total legal AI spend	$142K	The rebuild, one-time	$24K

THE RULE OF THUMB

Rebuild where the price is per seat and the work is an API call; keep what is a data source. **The case: $108K back in year one, the build paid back in under 10 weeks of saved Harvey billing, ~$132K a year from year two.** At 100 attorneys on Harvey's published pricing, year one saves ~$277K at today's fixed price; at 200, ~$552K.

Case figures from the AmLaw 200 case on our law-firms page; the 100- and 200-attorney figures from our Harvey rebuild table. Our fixed price starts at $16,600 for 20–50 attorneys (back page). If your own team builds it, put your engineering cost in place of the build.

Rebuild the wrappers. *Keep the data sources.*

A typical AmLaw 200 firm runs 6–10 vendors, and five of the ten are wrappers on commodity APIs. Not every line item should be replaced: the rebuild candidates are the wrappers, not the data sources.

REBUILD FIRST: THE LEGAL AI WORKFLOWS

- **Document review**: summaries, key issues, risk flags, cited answers
- **Drafting and redlining** from firm templates and matter context
- **Legal research** in plain language, against the databases you already subscribe to
- **Briefs, memos and opinions** in Bluebook, OSCOLA or AGLC format
- **Deposition and discovery summaries**: key facts, witness inconsistencies
- **Citation verification** against an authoritative source
- **Privilege walls and ethical screens** at matter level
- **DMS integration**: read-only or write-back
- **Billable-time attribution** and **per-matter cost** for client billing
- **Audit log** of every prompt and output

REBUILD NEXT, ON THE SAME MATTER MODEL

- **Matter and billing** (Clio, Filevine): a matter app on Postgres, QuickBooks or Sage behind it
- **Document storage** (NetDocuments, iManage): object storage, metadata and search, once the workspace is settled

KEEP AND CONNECT

- **Westlaw, Lexis, Bloomberg Law**: their APIs, through your enterprise contracts
- **Lex Machina** and other practice data feeds
- **Your DMS**, read in place until you choose to replace it

DROP ON PURPOSE

- **The vendor's fine-tuning corpus**: foundation models with your documents as context
- **Prompt libraries for 100+ tasks**: build the 10–20 your firm uses
- **Multi-firm benchmarking**: single-firm by design

Three layers. *All inside your tenancy.*

A Next.js app your IT team can security-review in one sitting. Documents stay in your encrypted bucket, prompts and outputs never leave your tenancy, and matter-level access controls match the walls you already run.

LAYER 1 · ATTORNEYS USE

The legal workspace

- Matter-scoped chat, draft and review
- Review panel: redlines, summaries
- Brief, contract and memo templates
- Citation-aware output
- Walls: partner, associate, paralegal
- Time-entry export to Clio or iManage

LAYER 2 · ROUTING

Model and retrieval

- Claude Sonnet 4.5 by default, GPT-4o for images
- Retrieval over matter documents, walls applied first
- A citation verification pass
- Token budget per matter and attorney
- An audit log of every prompt

LAYER 3 · YOUR ACCOUNTS

Contracts and sources

- **Anthropic** enterprise contract, yours
- **OpenAI** Platform API, yours
- Zero data retention in both contracts
- **Your DMS**: iManage, NetDocuments, SharePoint
- **Westlaw, Lexis** through your contracts

Decisions to make in week one

DECISION	OUR DEFAULT	WHY
Where it runs	Your AWS account, or on-prem	Documents never leave your tenancy
Sign-in	Your SSO from day one	Leavers and screens follow your directory
Documents	Connect the DMS; never migrate it	Your DMS stays the record
Search	pgvector in your own Postgres	Passages sit next to the access tables
Models	Two providers behind one adapter	Switch models by config, no retraining
Contracts	Zero retention, signed before ingest	Prompts are not kept or used for training

Compliance: on your own AWS account you inherit AWS's SOC 2 and ISO 27001 posture. Write the documentation against ABA Formal Opinion 512 (2024) and, for UK offices, SRA guidance as you build, not after.

The matter is the spine. *Everything hangs off it.*

What attorneys pay legal AI for is context: it knows which matter a question belongs to. Rebuild that as one matter record that every document, prompt, output and minute points to. Model it before any screen; matter management and billing reuse it later.

TABLE	KEY COLUMNS
`clients`	id, name, conflicts_id, ai_terms (from outside counsel guidelines), billing_terms
`matters`	id, client_id, dms_workspace_id, name, practice_area, jurisdiction, citation_format, status, billing_code, retention_policy, legal_hold
`users`, `matter_access`	id, email, role (partner, associate, paralegal); matter_id, user_id, access (full, read, none), source (DMS, conflicts)
`ethical_walls`	id, matter_id, screened_user_id, reason, set_by, effective_at, lifted_at
`documents`	id, matter_id, dms_id, dms_version, title, doc_type, privileged, sha256, synced_at
`passages`	id, document_id, matter_id, page, heading, text, embedding (vector)
`prompts`, `outputs`	matter_id, user_id, mode (chat, draft, review, research), template_id, model, passages_used, tokens_in, tokens_out, cost
`citations`	output_id, cited_text, format (Bluebook, OSCOLA, AGLC), source (Westlaw, Lexis, precedent DB), status (verified, flagged, removed)
`templates`	id, practice_area, doc_type, title, body, version, owner
`time_entries`	matter_id, user_id, minutes, ai_assisted, approved_by, exported_to

TWO RULES THAT SAVE A REBUILD

Filter by matter before the search, never after it: a passage the user may not open must never reach the model, because a draft cannot unsee it. **Put matter_id on every row**: cost by matter, retention by matter and legal holds by matter all fall out of that one column.

Six weeks, *pilot included.*

This is the order we build in: the walls before the documents, the documents before the drafts. Each week ends with something your practice lead can try, and the last is a pilot with five attorneys on live matters.

WEEK	BUILD	DONE WHEN
1	**Matters and walls.** Your AWS account or on-prem, your SSO, zero-retention terms signed with Anthropic and OpenAI. Clients, matters, access and ethical walls synced from the DMS and the conflicts system.	A screened user is blocked on a test matter
2	**Documents in place.** The DMS connector, read-only: open matters indexed inside your tenancy, split at headings, with DMS permissions inherited.	Search returns only what the user can open in the DMS
3	**The drafting workflow.** Chat, draft and review modes, the review panel, the firm's top templates as priors.	An associate drafts an engagement letter from a firm template in one sitting
4	**Research and citations.** Westlaw or Lexis through your contracts, the verification pass, formatting by jurisdiction.	Every citation in the test set is verified, flagged or removed
5	**Audit, cost, review.** The audit log with your retention windows, per-matter cost, time-entry export. Security review with IT and risk.	IT and risk sign off
6	**Pilot and rollout.** Five attorneys for a week on live matters with a daily fix list, then firm-wide rollout by practice group.	The pilot attorneys choose it over Harvey

THEN THE NEXT WRAPPER

Keep Harvey **until your renewal date**, then cancel; nothing in your DMS moved, so nothing moves back. With the workspace live, **matter management and billing** are the next rebuild: the matter model they need is already in place.

Privilege is the product. *Hold every wall.*

The models draft well out of the box. What a firm pays for is the boundary around them: documents that stay where your engagement letters promise, walls the model never sees past, and contracts that keep prompts out of anyone's training set.

1. **Documents never leave the tenancy.** Documents stay in your encrypted bucket, in your AWS region or on-prem, and the DMS stays the record. Index inside the tenancy and send a model only the passages a question needs.
2. **Sign zero retention before the first document.** Anthropic and OpenAI enterprise contracts with zero data retention: prompts and outputs are not used for training and are not retained server-side. No copy on any vendor's infrastructure, the builder's included.
3. **Enforce walls at retrieval.** Matter access and ethical screens filter the search before any passage reaches the model. Inherit DMS permissions on every sync, and make a new screen take effect at the next query, not the next re-index.
4. **Treat the log as privileged.** Prompts and outputs quote privileged material. Keep the audit log inside the tenancy, behind the same walls, with the retention windows your records policy sets and legal holds by matter.
5. **Honor each client's terms.** Some outside counsel guidelines restrict AI on their matters. Carry the terms on the client record and enforce them in the workspace, not in a memo.

Verify every citation

Resolve every cited case and statute against Westlaw, Lexis, the Caselaw Access Project or your precedent DB before output. A citation that cannot be verified is flagged or removed, never shown as settled. Format after the check, so clean formatting never dresses up a bad cite.

Answer the risk committee from the log

Every prompt, output, citation result and cost, by matter. It answers the ABA Formal Opinion 512 (2024) question of how AI was used on a matter, and it feeds the client bill.

Connect it, launch it, *run it.*

Connecting your stack

- ☐ List the legal-tech vendors: wrapper or data source
- ☐ List the workflows attorneys run in Harvey today, by practice group
- ☐ Pick the 10–20 prompt templates in real use
- ☐ Sign zero-retention terms with Anthropic and OpenAI before the first document
- ☐ Connect the DMS read-only; keep the DMS ID and version on every document
- ☐ Sync matter access and ethical walls from the DMS and conflicts system
- ☐ Record each client's AI terms on the client record
- ☐ Connect Westlaw or Lexis through your enterprise contracts
- ☐ Test the walls with a screened user before any attorney signs in

Go-live

- ☐ Attorneys signing in through your SSO
- ☐ Security review signed off by IT and risk
- ☐ Citation test set passed
- ☐ Audit-log retention set to your records policy
- ☐ ABA 512 and SRA documentation filed with risk
- ☐ Rollback ready: Harvey live until renewal

Running it

- ☐ **Always on:** uptime and error alerts, security patches, wall-sync alerts
- ☐ **Weekly:** flagged citations reviewed with the practice lead
- ☐ **Monthly:** model update tested against the citation set, template review
- ☐ **Quarterly:** token spend by matter against the bills you replaced, access review
- ☐ **Yearly:** a restore drill, a records-policy check, a wall audit

Who you need

A senior full-stack engineer at home with LLM APIs, retrieval and DMS APIs; a practice lead for the workflows and the five pilot attorneys; IT and risk for the walls and the records policy. Then one named owner: every query touches privileged documents, and someone answers when a wall sync fails.

Five ways a rebuild *stalls.*

1. Replacing the data sources too

Westlaw, Lexis, Bloomberg Law and Lex Machina are data, and data is worth keeping. Rebuild the wrapper on top and call their APIs; the subscriptions stay, the markup goes.

2. Swapping one wrapper for another

Spellbook and Legora are about 50% cheaper than Harvey today, on the same wrapper economics over the same APIs. Within 18 months their pricing rises and you are redoing this analysis.

3. Handing associates raw ChatGPT Enterprise

No matter context, no privilege walls, no DMS integration, no Bluebook citations. Associates stop using it, and the firm has paid for a chat window.

4. Self-hosting LibreChat and calling it done

LibreChat is a generic chat UI, not a legal workflow. Citations, retrieval, privilege walls and DMS integration are all still yours to build, often 2–3× our build cost in internal hours.

5. Migrating the DMS first

Moving every document before the AI workspace ships turns a six-week project into a records project. Connect the DMS in place; replace it later, on the matter model, if the numbers say so.

And one that is not a mistake: negotiating

Some firms negotiate Harvey down 15–25% on multi-year contracts. Do it if the rebuild is a year away; it captures about an eighth of what the rebuild does.

RATHER HAVE IT BUILT?

We build this exact plan. *Fixed price.*

The same matter model, walls and six weeks, delivered by a senior team that has shipped 250+ products.

$16,600 fixed for 20–50 attorneys: a senior engineer, a coder and a half-time delivery manager for 5 weeks; $22,100–$29,100 for 50–100 with multi-office tenancy; $33,200–$49,800 for 100+ with SOC 2 Type II posture, on-prem and multi-region

---

Live in 4–6 weeks, with a one-week pilot with five of your attorneys first

---

Documents never leave your tenancy: connected to your DMS and your Westlaw or Lexis contracts; nothing migrates

---

90-day warranty, and the code in your firm's GitHub from day one; we keep no copy

---

We can run it from $3,600 a year: hosting, monitoring and monthly model updates; usage bills to your own Anthropic and OpenAI accounts

alliedbiz.tech/rebuild/for-law-firms

Allied BizTech · 15+ years · ISO 9001:2015 · 250+ products shipped. Harvey, Spellbook, Casetext, Legora, Clio, Filevine, NetDocuments, iManage, SharePoint, Westlaw, Lexis, Bloomberg Law, Lex Machina, QuickBooks, Sage, ChatGPT, OpenAI, Anthropic, Claude, LibreChat and AWS are trademarks of their owners, named here for identification only; this guide is independent and not endorsed by any of them.