ALLIED BizTech SOLUTIONS

FIELD GUIDE · 2026

Rebuild Harvey yourself.
The plan we follow.

For firms of 20 to 200+ attorneys who want to own their document review, drafting and research workflows, keep every document in their own tenancy, and stop paying per attorney seat.

- → Is it worth it, at your size
- → What to rebuild and what to drop
- → The architecture, in three layers
- → Rollout and go-live checklists
- → The data model
- → A week-by-week build plan
- → Citations, privilege, residency
- → Where teams get stuck

Harvey charges per attorney. *The models charge per token.*

Harvey's per-seat enterprise pricing scales linearly with headcount. The work underneath runs on OpenAI and Anthropic models, and your firm's real usage is much lower than the seat price implies. A rebuild replaces the seat fee with tokens at your provider's price and a flat hosting bill.

FIRM	HARVEY A YEAR	REBUILT: YEAR ONE / AFTER	SAVED IN YEAR ONE
20 attorneys · ~100 documents each a month	~$66,000	~$21,100 / $4,500	~$44,900
50 attorneys · ~150 documents each a month	~$165,000	~$26,200 / $9,600	~$138,800
100 attorneys · ~200 documents each a month	~$330,000	~$53,100 / $24,000	~$276,900
200 attorneys · ~250 documents each a month	~$650,000+	~$97,800 / $48,000	~$552,200

Harvey enterprise pricing per public reporting (2024–2026 range). "Rebuilt" uses Anthropic Claude Sonnet 4.5 ($3 input / $15 output per 1M tokens) at 10–25K tokens a document, hosting on your AWS, and our fixed build price: $16,600 up to 50 attorneys, $29,100 (the top of $22,100–$29,100) to 100, $49,800 (the top of $33,200–$49,800) above 100. If your own team builds it, put your engineering cost in that place.

THE RULE OF THUMB

The math works from Harvey's 20-seat minimum: ~$44K saved in year one at 20 attorneys. **At 50 attorneys, the rebuild saves ~$138K in year one and ~$155K a year after.** From 100 attorneys, year one alone saves more than $270K.

Who you need

- **A senior full-stack engineer** at home with LLM APIs, retrieval (pgvector or Pinecone) and document-management APIs. Our team ships this in 4–6 weeks; one engineer working between other priorities should plan for longer.
- **A practice or knowledge lead**, two hours a week: the workflows that matter, the firm templates, the five pilot attorneys, the rollout call.
- **Your IT and risk team**: the security review, the privilege model and the records policy.
- **An owner after launch**: hosting, monitoring, the monthly model update and the flagged-citation review.

02 · SCOPE

Rebuild the 12 workflows your attorneys use. *Drop the rest.*

Harvey ships dozens of features and prompt libraries for 100+ legal tasks. A firm runs a defined subset every day, the 15–25% it pays for and uses. That subset is the project; the long tail is where rebuilds stall.

REBUILD ONE FOR ONE

- **Document review and summarisation**: summary, key issues, risk flags
- **Contract drafting and redlining** from template plus matter context
- **Legal research**: RAG over firm precedents plus the Westlaw or Lexis API
- **Brief, memo and opinion drafting** in the jurisdiction's citation format
- **Deposition and discovery summaries**: key facts, witness inconsistencies
- **Citation verification**: Bluebook, OSCOLA, AGLC
- **Privilege walls and ethical screens** at attorney and matter level
- **DMS integration** (iManage, NetDocs, SharePoint): read-only or write-back
- **Billable-time attribution**: AI-assisted hours logged by matter

REBUILD, SIMPLER

- **"Vault" workflows**: the 3–5 your firm uses
- **M&A diligence**: matched to your DD playbook
- **Prompt library**: the 10–20 templates in use

DROP ON PURPOSE

- **Harvey's fine-tuning corpus**: foundation models with your documents as context
- **Research databases**: keep Westlaw, Lexis or Bloomberg Law; call the API
- **Multi-firm benchmarking**: single-firm by design
- **Harvey Academy**: training recorded for your workflows

ADD WHAT PER-SEAT PRICING DISCOURAGES

- Per-matter cost reporting for client billing
- Audit-grade logging of every prompt and output
- New attorneys at no seat fee; models switch by config

Three layers. *Privilege-aware.*

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 access controls match the walls you already run.

LAYER 1 · ATTORNEYS USE

The legal workspace

- Matter-scoped chat and draft mode
- Review panel: redlines, summaries
- Brief, contract and memo templates
- Citation-aware output (Bluebook)
- 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
- RAG over the firm's documents
- A citation verification pass
- Token budget per matter and attorney
- An audit log of every prompt

LAYER 3 · BILLING

Your enterprise accounts

- **Anthropic** enterprise contract, yours
- **OpenAI** Platform API, yours
- Signed BAA and SOC 2 on both
- Zero data retention in the contract
- Per-matter cost reporting

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	The DMS, read-only first	Your DMS stays the record; nothing migrates
Search	pgvector in your Postgres	Passages sit next to the access tables
Model	Claude Sonnet 4.5, behind an adapter	Swap models without touching the workspace
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, HIPAA-eligible with a BAA. Write the documentation against ABA Formal Opinion 512 (2024) and, for UK offices, SRA guidance as you build, not after.

Ten tables *carry the whole product.*

Model these before writing any screen. Every row that holds firm content carries its matter_id, and every document keeps its DMS ID and version, so a re-sync can run twice and any answer traces back to its source file.

TABLE	KEY COLUMNS
`matters`	id, dms_workspace_id, client_id, name, practice_area, jurisdiction, status, billing_code, opened_at
`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, ingested_at
`passages`	id, document_id, matter_id, page, heading, text, embedding (vector)
`prompts`	id, matter_id, user_id, mode (chat, draft, review, research), template_id, body, created_at
`outputs`	id, prompt_id, matter_id, model, passages_used (jsonb), body, tokens_in, tokens_out, cost, created_at
`citations`	id, output_id, cited_text, normalised_cite, format (Bluebook, OSCOLA, AGLC), source (Westlaw, Lexis, Caselaw Access Project, precedent DB), status (verified, flagged, removed), checked_at
`templates`	id, practice_area, doc_type, title, body, version, owner
`time_entries`	id, matter_id, user_id, minutes, ai_assisted, approved_by, exported_to (Clio, iManage)

TWO RULES THAT SAVE A REBUILD

Filter by matter before the search, not after it: a passage the user may not open must never reach the model, because a draft cannot unsee it. **Write every prompt and output to an append-only log** with matter, model, tokens and cost: it is your audit trail, your records store and your client-billing report in one.

Six weeks, *pilot included.*

This is the order we build in. 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	**Tenancy and walls.** Your AWS account or on-prem, your SSO, zero-retention terms signed on your Anthropic and OpenAI accounts. Matters, users, access and ethical walls synced from the DMS and conflicts system.	A screened user is blocked on a test matter
2	**Documents and retrieval.** The DMS connector, read-only: open matters ingested, split at headings, embedded into pgvector with matter_id and the DMS permissions inherited.	A search returns only what the user can open in the DMS
3	**The workspace.** Chat, draft and review modes; the review panel with redlines and summaries; the firm's top templates.	An associate drafts an NDA from a firm template in one sitting
4	**Research and citations.** Westlaw or Lexis through your enterprise contracts, the citation verification pass, Bluebook, OSCOLA and AGLC formatting (next page).	Every citation in the test set is verified, flagged or removed
5	**Logging, cost, review.** The audit log, per-matter cost report, token budgets, time-entry export to Clio or iManage. Security review with your IT team; SOC 2 and ABA 512 documentation.	IT signs off the security review
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

AFTER ROLLOUT

Keep Harvey **until your renewal date**, then cancel. Nothing in your DMS moved, so there is nothing to move back, and the partners who want a second week of overlap get it.

Check every citation. *Hold every wall.*

The models draft well out of the box. What a firm pays for is the checking around them: citations that exist and say what the draft claims, privilege walls the model never sees past, and documents that stay where your engagement letters promise.

1. **Verify before output.** Pull every cited case and statute out of the draft and resolve it against an authoritative source: Westlaw, Lexis, the Caselaw Access Project or your precedent DB. Confirm the pin cite and the passage it supports. A citation that cannot be verified is flagged or removed, never shown as settled.
2. **Format after you verify.** Normalise to Bluebook, OSCOLA or AGLC from the matter's jurisdiction once the check has passed, so clean formatting never dresses up a bad cite.
3. **Test on your own work.** Build a test set from past memos and briefs with known-good citations, then plant fakes and near-misses. Launch when every fake is caught and the practice lead signs off; rerun the set on every model update.
4. **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.
5. **Keep data resident.** Documents in your encrypted bucket, in your AWS region or on-prem. Zero-retention enterprise contracts with Anthropic and OpenAI, so prompts and outputs are not used for training or retained server-side. No copy on any vendor's infrastructure, ours included.

Show the reviewer the check

Mark each citation in the draft as verified or flagged, with its source one click away. The reviewing attorney sees what was checked and spends the review on judgement, not on retrieval.

Log for the auditor and the client

Every prompt, output, citation result and cost, by matter, retained per your records policy. It answers the ABA Formal Opinion 512 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 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
- ☐ Connect Westlaw or Lexis through your enterprise contracts
- ☐ Load firm templates and precedents for drafting
- ☐ 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
- ☐ 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:** cost review, token spend by matter against the Harvey bill you replaced, access review
- ☐ **Yearly:** a restore drill, a records-policy check and a wall audit

Who owns it

Name one person before launch. Every query touches privileged documents: someone has to answer when a wall sync fails or a model update changes the drafts.

Five ways a rebuild *stalls.*

1. 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.

2. Taking the research bundle

Casetext, now Westlaw AI, is useful if you already pay for Westlaw. It sits on GPT-4 with the same per-seat structure, so the wrapper margin stays.

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. Skipping the pilot

Five attorneys for a week find what a test plan misses: the template nobody listed, the screen a sync missed, the partner who wants redlines rather than summaries.

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 scope, architecture 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

---

Connected, not migrated: iManage, NetDocs or SharePoint, and your Westlaw or Lexis contracts

---

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

---

We can run it for $3,600 a year: hosting, monitoring and monthly model updates

alliedbiz.tech/rebuild/harvey-alternative

Allied BizTech · 15+ years · ISO 9001:2015 · 250+ products shipped. Harvey is a trademark of its owner, named here for identification only; this guide is independent and not endorsed by Harvey. Harvey pricing per public reporting, 2024–2026.