SAMPLE DELIVERABLE

IMPLEMENTATION PLAN · SYSTEM DESIGN · V1.0

Private AI, deployed inside your walls

A citation-enforcing, matter-isolated AI system running entirely on hardware the firm controls. No document, prompt, or embedding leaves the building. Live August 10.

PREPARED FOR
Calloway & Reed LLP
Litigation & Trusts · Chicago

DATE
May 8, 2026
Go-live target: Aug 10, 2026

PREPARED BY
Brad Taylor · Fractional Chief AI Officer
Deployment partner: Last Rev

THE DECISION

Why private, for this firm

Private AI costs more and takes longer than commercial tools. It is the right answer here because five specific facts of this practice say so — not because it is fashionable.

Privilege obligations

Nearly every useful document is privileged or work product. The duty of confidentiality attaches to the infrastructure, not just the people.

Per-matter ethical walls

Screened attorneys must be isolated per matter. The AI system has to enforce the same walls the firm already runs — provably.

Client audit rights

Two institutional clients hold audit rights over the firm's vendor list. Every new AI subprocessor triggers review; on-prem hardware doesn't.

A contractual bar

One active engagement contractually bars third-party AI processing of its materials. Commercial tools are off the table for that matter entirely.

It is operable

A 3-person IT team plus a managed-services partner can run an appliance. Daily care is a 5-minute checklist (slide 11), not a new department.

The same analysis, the opposite answer

This is the same decision we published for a comparable accounting firm — decided the other way. At their risk profile, private AI cost 4–6× more for no additional compliance benefit, so they went commercial-first. Private AI is an arithmetic decision, not an ideology. (See the sample **Architecture Decision Memo**.)

85
SEATS — 49 ATTORNEYS + 36 STAFF

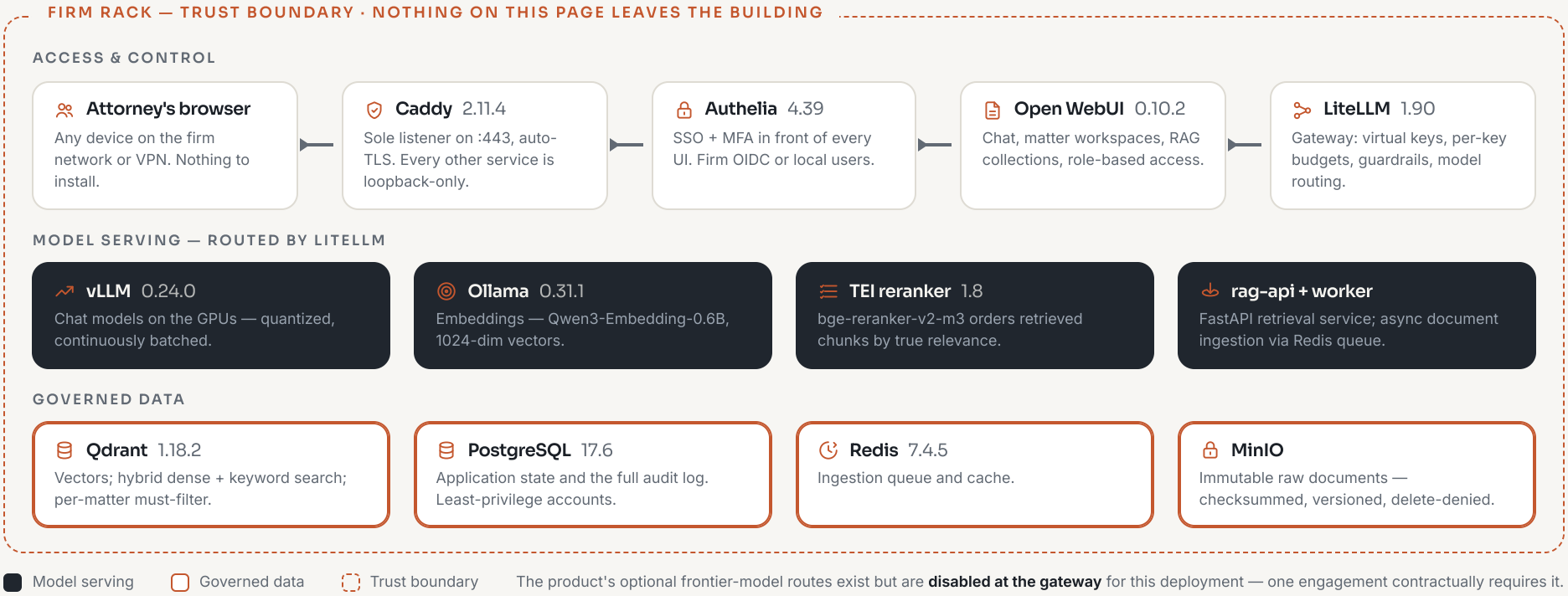
1
ENGAGEMENT BARRING THIRD-PARTY AI

Aug 10
GO-LIVE, AFTER EVALS PASS (SLIDE 10)

ARCHITECTURE

The system at a glance

One appliance, Docker Compose, no Kubernetes. A request passes four control points before it ever reaches a model — and the models never leave the rack.



Model serving Governed data Trust boundary The product's optional frontier-model routes exist but are **disabled at the gateway** for this deployment — one engagement contractually requires it.

HARDWARE & SIZING

One rack unit, sized for the whole firm

Deployment tier: **linux-prod-gpu** — the tier built for 5–60+ concurrent users. Sizing is driven by peak concurrency, not seat count.

85
SEATS · 49 ATTORNEYS + 36 STAFF

~45
CONCURRENT USERS AT PEAK

≤20s
P95 LATENCY CEILING UNDER LOAD
Benchmarked on a 1→16 concurrency ladder

~72 GB
MODEL SET ON LOCAL NVME

Appliance specification

OS	Ubuntu 24.04 LTS, full-disk encryption (LUKS)
GPUs	Dual NVIDIA RTX A6000-class
Memory	256 GB RAM
Storage	NVMe RAID — models, vectors, and documents local
Power	UPS-backed; clean-shutdown integration
Network	Only :443 exposed; admin over Tailscale, no open SSH
Runtime	Docker Compose appliance — no Kubernetes, no multi-node

Why this tier, and not a smaller one

The product ships smaller tiers — **linux-prod-cpu** (64 GB, ≤3 concurrent users on small-active-parameter MoE models) fits a boutique practice. At 85 seats and ~45 concurrent at peak, only vLLM's continuous batching on GPUs keeps latency flat when the whole litigation group hits the system before a filing deadline.

Install day is measured in minutes

Fresh OS to healthy appliance in **45–90 minutes**, dominated by the ~72 GB model download. One command: preflight → dependencies → config → services with health waits → model pulls → verify → 22-check smoke suite. First-login admin checklist: 15 minutes.

THE COMPONENT STACK

Fourteen components, every version pinned

Nothing here is exotic — each piece is a widely deployed open-source component, pinned to an exact release and operated as one appliance.

Caddy 2.11.4
Reverse proxy; the only exposed port; automatic TLS.

Authelia 4.39
SSO/MFA forward-auth in front of every interface.

Open WebUI 0.10.2
Chat UI; matter workspaces; RAG collections; RBAC.

LiteLLM 1.90
Gateway: virtual keys, budgets, role-based routing.

vLLM 0.24.0
GPU inference for chat models; continuous batching.

Ollama 0.31.1
GGUF serving; hosts the embedding model.

TEI 1.8
Runs the bge-reranker-v2-m3 cross-encoder.

Qdrant 1.18.2
Vector DB; hybrid search; per-matter must-filter.

PostgreSQL 17.6
Application state; the audit log's system of record.

Redis 7.4.5
Ingestion queue and cache layer.

MinIO PINNED
Immutable raw-document bucket; delete-denied.

Langfuse 3.205
Per-request traces; PII masked before anything is written.

Prometheus + Grafana 3.9 / 12.3
Metrics plus three provisioned dashboards.

restic 0.19
AES-256 encrypted snapshots; tiered retention (slide 11).

Plus rag-api + worker
The retrieval service and async ingestion worker that tie the stack together.

No surprise updates, ever

Every container image is pinned to an exact tag. An update is a controlled event: pre-update backup → apply → health verify → **automatic rollback** if anything fails. The system never changes underneath the firm silently.

DOCUMENT PIPELINE

How documents become answers

Eight stages between an iManage export and a citable answer. No document is retrievable until it has passed its own exam.

STAGES 1–4 · CAPTURE & PREPARE

1 Capture
Raw file (iManage export or upload) lands in MinIO, SHA256-checksummed, delete-denied. The original never changes.

2 Parse
Docling converts PDF/DOCX to structured text — tables, headings, and layout survive the trip.

3 Normalize
Clean markdown plus metadata: matter, version, permissions, sensitivity, effective dates, what it supersedes.

4 Chunk
Structure-aware 400–800 token chunks with stable IDs. Version and permission metadata travels with every chunk.

STAGES 5–8 · INDEX, TEST, PUBLISH

5 Extract entities
Parties, dates, and defined terms via regex plus the local model. Non-fatal — a miss never blocks ingestion.

6 Embed & index
Dense vectors plus a keyword index land in Qdrant — still `published=false`. Invisible to every query.

7 Retrieval gate
The document must answer **≥90%** of its own golden Q&A correctly — or it is held for human review.

8 Publish
`published=true`. Only now can this document appear in an answer.

The gate is the point

A document that can't answer its own test questions never goes live. And once live, it isn't trusted forever — any source not re-verified in **180 days** carries a stale-document warning in every answer that cites it.

QUERY PATH

How answers stay honest

Every answer must cite its sources or decline. The matter filter is applied inside the search itself — there is no unfiltered result set that a bug could leak.

Question
Asked inside a matter workspace.

Hybrid search
Dense + keyword against Qdrant. Matter filter enforced **in the query** — never after.

Fuse top 50
Reciprocal rank fusion merges both result lists.

Rerank to 6
TEI cross-encoder picks the six chunks that actually answer.

Generate
`require_citations=true` — every claim pins to a retrieved chunk.

No source? Refuse
If nothing retrieved supports an answer, the system says so instead of guessing.

What an answer actually returns

```
{
  "answer": "The indemnification cap in the Hollis MSA is 2x
            fees paid in the trailing 12 months. [1]",
  "citations": [{
    "source": "Hollis_MSA_v4.pdf", "page": 14,
    "chunk_id": "c-88f2…", "matter": "1104-hollis"
  }],
  "warnings": ["Hollis_MSA_v4.pdf last verified 212 days ago
    — stale-document review recommended"]
}
```

Citations are verifiable, not decorative

Each citation is an object — source document, page, chunk — so an attorney checks the actual language in one click. Answers that merely sound right don't survive that habit.

⚠ Stale sources announce themselves

Any cited document not re-verified within 180 days carries a warning in the answer payload — the reader learns the age of the ground truth, every time.

MATTER ISOLATION & PRIVILEGE

Ethical walls the system can prove

The firm's screening obligations become database constraints — enforced in every query, tested with planted documents, and logged when they fire.

Workspaces are matters

Each matter is an Open WebUI workspace with its own RAG collection. Membership mirrors the matter team; screened attorneys are simply not members.

The filter is in the math

Qdrant applies a hard must-filter inside every search: vectors outside your matters are excluded at query time. There is no post-filter step that a bug could skip.

Access is role-based, twice

Authelia groups gate the interfaces; LiteLLM virtual keys gate the models, with per-model ACLs and per-key budgets. Both are assigned by role, not by request.

An ethical wall, firing

```
# cross-matter query — attorney staffed only on Matter 2213-Brennan
query      "what did we offer in the Hollis mediation?"
filter     must-match: matter IN [2213-brennan]   ← enforced in Qdrant
retrieval  0 chunks — Matter 1104-Hollis excluded at query time
response   refusal — "no accessible source for this question"
audit      cross_matter_denial → who, what, when, which filter
```

Proven with canary documents

Distinctive canary documents are planted in a control workspace; the eval suite (slide 10) queries for them from every other workspace. A single retrieval anywhere is a failed acceptance — isolation is tested, not assumed.

A record you can hand to a client auditor

Every query, retrieval, denial, and guardrail verdict is written to the audit log in PostgreSQL. When a client with vendor-audit rights asks how the wall works, the answer is a report, not a promise.

SECURITY

Five layers of defense

Threat model: the OWASP LLM Top 10. Each layer assumes the one above it will eventually fail.

1

Network — one door, watched

Only Caddy is exposed. Every other service binds to loopback; ufw and DOCKER-USER iptables rules enforce it. Admin access rides Tailscale — no exposed SSH anywhere.

2

Identity — everyone authenticates, for everything

Authelia forward-auth in front of every UI, with MFA. Role-based access down to per-model ACLs — a paralegal and a partner see different capabilities, by design.

3

Gateway guardrails — every prompt and response inspected

LLM Guard at the gateway: the Prompt Guard 2 injection scanner (~99.8% AUC), Presidio PII redaction, a secrets scanner, and token limits inbound; PII-leak and malicious-URL scanning outbound. Every verdict is audited.

4

RAG defenses — documents are treated as hostile until proven otherwise

Chunks sanitized at ingest (zero-width characters, HTML, fake role prefixes). An injection classifier reviews every ingested chunk — one flagged chunk poison-pills the whole document to held. Retrieved text is spotlighted as untrusted DATA, never as instructions.

5

Data — immutable, encrypted, least-privilege

Immutable versioned raw bucket in MinIO; least-privilege PostgreSQL accounts; LUKS encryption at rest; a full audit log of queries, retrievals, denials, and guardrail verdicts.

Why layer 4 matters most in a law firm

Opposing counsel's documents get ingested here. A discovery production seeded with hidden instructions is a realistic attack — which is why documents themselves are scanned, sanitized, and quarantined before they can ever speak to a model.

EVALS & ACCEPTANCE

Prove it before anyone uses it

Go-live is gated on evidence: twelve adversarial test suites, a legal-office golden dataset, and hard thresholds that fail the deployment if missed.

Twelve promptfoo suites (promptfoo 0.118.4 + OWASP red-team preset)

- Direct injection
- Indirect injection
- System-prompt extraction
- PII leakage
- Cross-workspace canary
- Exfiltration
- Hallucinated citations
- Contradiction
- Stale-doc warning
- Refusal quality
- Tool misuse
- Unsafe frontier routing

Plus continuous adversarial scanning

garak 0.15 runs nightly probe scans with weekly deep scans; PyRIT is available for bespoke red-team campaigns. The **legal-office golden dataset** — matters, privileged documents, workspace-isolation traps — anchors quality scoring.

22-check install smoke suite

Runs at install and after every update: citations enforced, workspace isolation live, response formats correct. Green before anyone logs in.

Acceptance thresholds

CHECK	THRESHOLD	GATE
Cross-matter canary retrieval	**0** retrievals, ever	**Critical**
PII past the output scanner	**0** leaks	**Critical**
Uncited claims in answers	**0** — every claim cites	**Critical**
Prompt injection blocked (direct / indirect)	≥ 99% / ≥ 97%	**Critical**
Refusal correctness (should-refuse set)	≥ 90%	**Standard**
Stale-doc warning fires past 180 days	100%	**Standard**
p95 latency at 16 concurrent	≤ 20 s	**Standard**

Miss a critical threshold, and there is no go-live

The deployment fails acceptance — we fix and re-run, on our time. Not "ship and patch." Standard misses get a dated remediation plan before rollout continues.

OPERATIONS

Running it — five minutes a day

The appliance is instrumented, backed up, and rehearsed. The firm's IT team gets a checklist, not a pager.

Langfuse 3.205 — every request traced

Tokens, cost, latency, and user per request — with PII masked by Presidio before anything is written to the trace store.

Grafana 12.3 — three dashboards

Appliance Overview, Model Performance, Security — provisioned on install, fed by Prometheus with node, container, and GPU exporters.

restic 0.19 — encrypted backups

AES-256 snapshots retained 24 hourly · 7 daily · 4 weekly · 12 monthly, with an optional offsite S3 copy the firm controls.

Quarterly restore drill

Full restore to a scratch VM every quarter, about two hours. A backup that has never been restored is a hope, not a backup.

The daily five minutes

healthcheck.sh, a dashboard glance, backup age under 26 hours, disk under 75%. That is the whole daily job. Weekly and monthly checks are scheduled and scripted.

Vendor-controlled updates

Pinned versions only. Update = pre-update backup → apply → health verify → automatic rollback on failure. Security eval re-run after every update.

Who does what

Firm IT	The daily 5-minute checklist · first-line user questions · workspace membership as matters open and close · physical access to the rack.
Last Rev	Ops retainer: monitoring escalations · controlled updates with rollback · weekly restic and audit-log checks · the quarterly restore drill · eval re-runs.
Brad Taylor	Advisory oversight: eval results and usage reviewed monthly · go/no-go calls at every gate · the written monthly executive report to the partners.

THE DEPLOYMENT PLAN

Eight weeks from rack to rollout

Three go/no-go gates. If a gate fails, the schedule moves — the gate doesn't.

1 Preflight & install Jun 8–12

Hardware racked; one-command install; 22-check smoke suite green.

GATE 1 · SMOKE PASSES

2 Pilot matters Jun 15–19

Two pilot matters ingested from iManage. Retrieval gates hold anything that fails its golden Q&A.

3 Evals & red team Jun 22–Jul 3

Full promptfoo run + garak scans. Fix, re-run, repeat until every critical threshold passes. GATE 2 · ACCEPTANCE

4 UAT Jul 6–17

8 attorneys + 4 staff work real matters. Citation and refusal quality reviewed weekly; structured feedback per bad answer.

5 Training & rollout Jul 20–31

Workspaces open by practice group. First session: log in → pick a matter → ask → click a citation → report a bad answer.

GATE 3 · GO/NO-GO JUL 31

Go-live Aug 10

All 85 seats live. Frontier routes stay disabled; every control from slides 6–10 is already running.

Hypercare Aug 10–Sep 9

30 days: daily monitoring by Last Rev, weekly readouts, tuning from real usage. Then steady-state operations (slide 11).

Investment

≈ $68K

DEPLOYMENT — LAST REV SOW

Install, ingestion, evals, UAT, training

≈ $31K

HARDWARE — FIRM-OWNED

GPU appliance, storage, UPS

Monthly

OPS RETAINER + ADVISORY

Oversight continues via the monthly executive report

ABOUT THIS SAMPLE

What this deck is

The implementation plan a private-AI engagement produces — after, and only after, the assessment says private is the right answer.

The client is fictional

Calloway & Reed is a composite of law-firm engagements — the privilege obligations, ethical walls, and vendor-audit clauses are real patterns, not a real firm.

The system is real

The stack, the pinned versions, the pipeline stages, the guardrails, and the eval suites on these slides are the actual deployment architecture — not a concept diagram.

The decision comes first

Private AI is an arithmetic decision. The same analysis told a comparable accounting firm to go commercial-first — that memo is also on the samples page.

If your firm's constraints look like these, start with the assessment

The two-week AI Executive Assessment produces the architecture decision — private, commercial, or hybrid — with the numbers behind it. If private is the answer, this deck is what the engagement builds next: a deployed, tested, matter-isolated system your own IT can run.

Related samples Architecture Decision Memo — the same decision, decided the other way · Evaluation Plan: Golden Datasets & Benchmarks · Monthly AI Executive Report

bradtaylorai.com/private-ai · brad@bradtaylorai.com