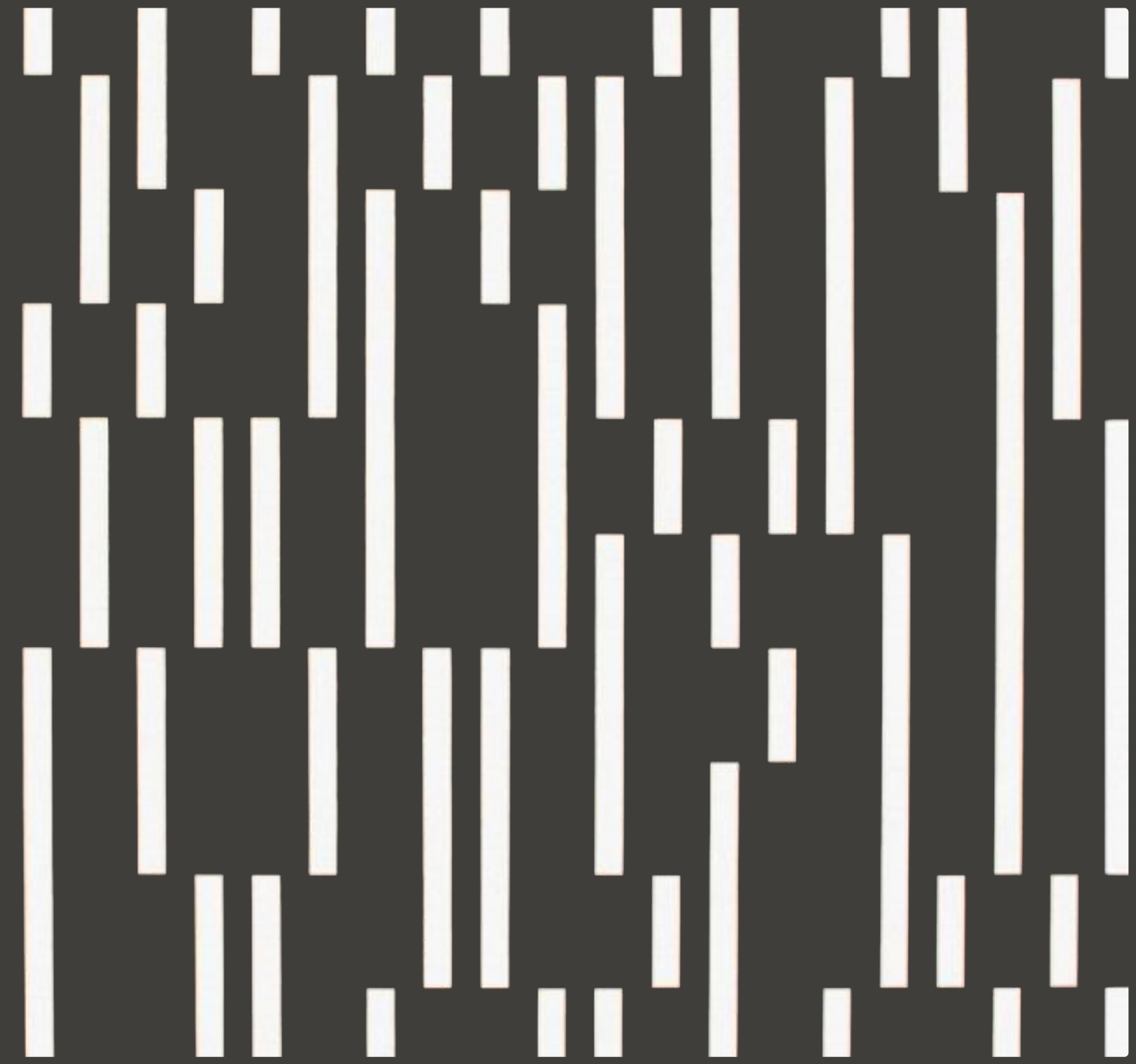


August

Planning Guide for *Legal AI* in 2026

A practical framework for evaluating, selecting and budgeting legal AI tools.

The Legal AI Adoption Gap

Every law firm has heard the pitch. AI will review contracts faster, summarize documents instantly, draft memos in seconds. The demos look great. The vendors are convincing. And yet, six months after signing, most firms find the tool barely used.

This pattern is predictable. A firm rolls out a new AI tool to a small group. The first week goes well. People explore, test a few workflows, see some promise. By the end of the month, usage drops. The early excitement fades. Skeptics feel proven right. The tool becomes another forgotten login.

The problem is not that the technology fails. In isolation, most tools do what they claim. The real issue is fit. They do not match how lawyers actually work.

Legal work lives in Word, Outlook, and the firm's document system. Now imagine a tool that forces you to leave all that, open a separate site, upload files, wait, then paste results back into Word. That does not save time. It adds friction.

Then there is trust. When a tool gives an answer without clear citations or links to the source, verification becomes manual. Do that a few times and you stop using it.

Finally, the output is often generic. The AI does not know your firm's style, your client's red lines, or the fallback language you have negotiated for years. The result may be correct, but it is not usable. You end up rewriting it anyway.

Firms that ask tough questions are not resisting change. They are responding to experience.

Executive Summary

AI has moved from novelty to necessity for competitive law firms. But the gap between a polished demo and defensible, day-to-day value remains significant. This guide provides a repeatable evaluation framework that firm leaders can use to compare tools, build a business case, and make confident budget decisions for 2026.

The framework maps each evaluation area to the outcomes that matter: quality, speed, risk reduction, and economics. It explains what to look for, what evidence to collect, and what red flags should give you pause. Whether you're running a pilot, conducting a bake-off, or reviewing an existing tool for renewal, this guide gives you the structure to make decisions any partner can trust.

Why This Guide Exists

The legal AI market is noisy.

Dozens of vendors now promise to handle everything from contract review to litigation support, and the demos can be impressive. But impressive demos don't pay for themselves. Firms that have been through multiple pilots know the pattern: a tool looks great in a controlled presentation, struggles with real documents, and ends up gathering dust after the initial enthusiasm fades.

This guide exists because evaluating legal AI requires a different approach than evaluating traditional legal technology. The outputs are probabilistic, not deterministic. Accuracy varies by document type and complexity. And the real test isn't whether the tool can do something once in a demo; it's whether lawyers will actually use it on day 30, day 60, and day 180.

We've built this framework based on what we've learned from working with law firms directly: what questions actually predict success, what evidence matters for procurement decisions, and what red flags should stop an evaluation early. The goal is to help you move from "this looks promising" to “here's the data that supports our decision.”

Where Legal AI Gets Hard

The early use cases for legal AI were relatively simple: “summarize this agreement”, answer a question about a contract, “give me a first draft of a memo.” Most tools can do these things reasonably well, at least on straightforward documents.

The real test comes when you try to use AI as part of your regular practice, not as an occasional experiment but as something your lawyers rely on for actual client work.

That's when you start running into harder questions.

Does it work on complicated documents? A tool that handles a short NDA well might struggle with a 200-page credit agreement full of cross-references and unusual formatting. What you see in a vendor presentation rarely reflects what you'll encounter on a live deal.

Can lawyers use it without changing how they work? If the tool requires your team to learn a new system, upload documents manually, and toggle between browser tabs and Word, most of them won't bother. People are busy. They'll stick with what they know.

Does it understand how your firm does things? Every firm has its own preferences. Certain ways of phrasing limitation of liability clauses. Particular positions on termination rights. Language you've refined over hundreds of deals. A tool that ignores all of that and gives you generic output isn't helping.

Can you trust what it tells you? This is the big one. If a lawyer can't quickly verify where an answer came from, they're not going to put it in front of a client. They'll either double-check everything manually, which erases the time savings, or they'll stop using the tool entirely.

Does it hold up under real volume? Running the tool on five documents during an evaluation doesn't tell you what happens when a whole practice group is

using it on a tight deadline. Some tools slow down. Some start making mistakes. You don't find out until it's too late.

Is the vendor going to be around when you need them? Legal AI is still early. The tools are evolving quickly, and you'll need help along the way. A vendor who takes three days to respond to a question during your evaluation will take even longer once you've signed.

These aren't theoretical concerns. They're the reasons partners push back on AI initiatives and why procurement asks uncomfortable questions. The skepticism is earned.

Why Legal AI Evaluations Go Wrong

If you've talked to other firms about their experience with legal AI, you've probably heard some version of the same story: we tried it, it didn't stick, we're not sure we want to try again.

Usually there's a pattern behind what went wrong.

The evaluation measured the wrong things. Procurement teams often work from a checklist: Does the tool summarize documents? Does it have a Word add-in? A tool can check every box and still fail in practice. The Word add-in might be clunky. The document management integration might require manual uploads. The summarization might fall apart on anything longer than ten pages.

The demo didn't reflect reality. Vendors are good at presentations. They choose documents that make the tool look good. They avoid edge cases. They show the product working exactly as designed. Then your lawyers try it on the messy, complicated, non-standard documents that fill their actual workload, and the results aren't nearly as impressive.

The tool handled tasks but not workflows. Being able to summarize a document or extract a few key terms is useful. But if every task requires starting from scratch, the time savings stay small. The lawyers who benefit most from AI are the ones who can build repeatable processes: upload these documents, extract these fields, flag these issues, generate this output. If the tool can't do that, the gains don't compound.

Features were promised but not delivered. It's common for vendors to sell capabilities that are "coming soon." Playbooks. Better integrations. Faster processing. Firms sign contracts expecting those features to arrive and then wait months for roadmaps that keep slipping. By the time the feature ships, if it ever does, the initial enthusiasm is gone.

The integration was more limited than it appeared. A tool might advertise Word and Outlook integration, but that can mean a lot of things. Sometimes it's a full-featured add-in that works seamlessly. Sometimes it's a basic plugin that pastes text and breaks your formatting. The difference between those two experiences is the difference between adoption and abandonment.

Accuracy problems destroyed trust. If the tool gives wrong answers, even occasionally, lawyers stop trusting it. That's true even when most of the outputs are correct. A few bad experiences are enough to poison the well. And once a tool has a reputation for being unreliable, it's very hard to get people to try it again.

These aren't obscure edge cases. They're the most common reasons legal AI doesn't stick. The rest of this guide is designed to help you avoid them.

How to Use This Guide

This guide is designed for a 2-4 week evaluation window.

Here's the recommended approach:

1. **Run a pilot** on real work product with participants across relevant practice groups.
2. **Capture structured feedback** using the evaluation questionnaire provided with this guide for each tool tested.
3. **Score with the rubric** and convert findings into a clear Adopt / Proceed with Guardrails / Do Not Adopt recommendation.
4. **Archive the evidence** (documents, screenshots, timing data) to support procurement, security review, and client conversations.

Scoring Framework

A structured scoring approach serves two purposes.

First, it forces evaluators to move beyond gut reactions and document-specific observations. Second, it creates a defensible record for leadership and procurement. Partners reviewing your recommendation shouldn't have to wonder how you reached your conclusion; the scores and supporting evidence should make the rationale clear.

Use a 1-5 scale across all evaluation areas:

Score	Awareness
1-2	Below threshold. Creates risk or need for rework. Do not proceed.
3	Adequate for limited use with guardrails and oversight.
4	Strong fit with clear, measurable gains.
5	Best-in-class. Strategic differentiator for the firm.

Evaluation Categories

We've consolidated evaluation into five categories that capture what actually matters for adoption and ROI.

Adjust weights based on your practice mix and priorities:

Category	Weight	Your Weight
Accuracy & Reliability	30%	
Firm Knowledge & Precedent Integration	25%	
Platform Integration & Workflow Fit	20%	
Advanced Capabilities	15%	
Vendor Partnership & ROI	10%	

QUICK REFERENCE

Five Questions That Matter

Use this as a standalone checklist or share it with pilot participants before evaluation begins.

- [ ] **Is it accurate enough to trust?**
  Test on real documents. Track errors by type and severity. If accuracy drops significantly on complex or non-standard documents, that's your answer.

- [ ] **Does it actually use our precedents and knowledge?**
  Generic AI is table stakes. Value comes from tools that learn your firm's style, leverage your document libraries, and apply your playbooks consistently.

- [ ] **Does it work where lawyers already work?**
  If lawyers have to leave Word or Outlook to use it, they won't. Look for native integrations that preserve context across applications, not bolt-on plugins.

- [ ] **Can it handle the hard stuff?**
  Basic summarization is solved. The real test is structured extraction at scale: tabular reviews across hundreds of contracts, deviation analysis against playbooks, and comparison matrices.

- [ ] **Is the vendor a partner or just a provider?**
  How they engage during the pilot predicts how they'll engage after you sign. Slow responses, generic training, and ignored feedback are leading indicators.

KEY CATEGORY 1

Accuracy & Reliability (30%)

The question: Can you trust it when it matters?

Why It Matters

Accuracy errors impact client work, court filings, and reputations. Usability or efficiency gains mean little if outputs are incorrect or misleading. This is a critical threshold. A tool that's 95% accurate on simple NDAs but only 70% on complex credit agreements isn't ready for serious tasks.

What Good Looks Like

- Measurably low hallucination rate with clear refusal behavior when the tool is uncertain
- Source attribution by default (document anchors, citations, authority links)
- Consistent results on adversarial or edge-case prompts
- Accuracy holds across document types, complexity levels, and non-standard formats

Evidence to Collect

Before the pilot: Have users define what "accuracy" means for their use case. Prepare gold-standard test documents with known answers.

During the pilot: Keep a separate accuracy log tracking each identified error with type, severity, and cause. Spot-check 10-15 outputs for citation accuracy.

After the pilot: Calculate precision, recall, and overall accuracy rate. Distinguish systemic versus isolated errors.

Red Flags

- "Sounds right" output without citations to source documents
- Fabricated citations or hallucinated authorities
- Consistent minor factual errors (wrong dates, swapped parties)
- Vendor claims "100% accuracy"

KEY CATEGORY 2

Firm Knowledge & Precedent Integration (25%)

The question: Does it leverage what makes your firm valuable?

Why It Matters

Generic AI capabilities are increasingly commoditized. The differentiated value comes from tools that can access and apply your firm's accumulated knowledge: precedent documents, clause libraries, client playbooks, and institutional style. A tool that generates generic boilerplate when you have years of refined precedents is a missed opportunity.

What Good Looks Like

- Direct integration with your document management system and precedent libraries
- Ability to apply client-specific or practice-specific playbooks consistently
- Drafting that matches your firm's style and terminology
- Configurable review standards that reflect your risk tolerance

Red Flags

- Outputs use generic boilerplate when firm precedents exist
- No clear path to integrate with your DMS
- Playbook functionality is promised but not yet built

KEY CATEGORY 3

Platform Integration & Workflow Fit (20%)

The question: Does it work where lawyers already work?

Why It Matters

Lawyers work in Word, Outlook, and their document management systems. They're not going to switch to a separate browser tab for every task, and they shouldn't have to. The best AI capabilities are useless if accessing them requires breaking out of existing workflows.

The Integration Question

Not all integrations are equal. There’s a real difference between a basic plugin that sends text to an API and a native integration that preserves context, formatting, and workflow across tools. Look beyond the checkbox. Can you start in Outlook, continue in Word, and revisit it in the web app without losing context? Can it access your DMS directly? Does the Word plugin preserve formatting and track changes?

What Good Looks Like

- Native integrations with Word and Outlook that feel like part of the application
- Context preservation across platforms
- Direct DMS integration without manual file uploads
- Less than 60 minutes to basic productivity for new users

Red Flags

- "Integration" means a basic plugin with limited functionality
- Context is lost when moving between applications
- More than 3 clicks to perform common actions

KEY CATEGORY 4

Advanced Capabilities (15%)

The question: Can it handle the work that actually moves the needle?

Why It Matters

Basic summarization and simple Q&A are table stakes. Every tool on the market can summarize a document. The differentiated value comes from capabilities that address tasks requiring significant associate time: structured extraction across large document sets, systematic comparison against standards, and consistent application of review criteria at scale.

What Good Looks Like

Tabular review: Extract specific data points across hundreds of contracts into structured formats

Playbook-based analysis: Systematically flag deviations from your standard terms

Comparison matrices: Generate side-by-side analysis of key terms across document sets

Scale handling: Process large document volumes without degradation

Red Flags

- Demo focuses on summarization without showing structured extraction
- Platform slows down or errors with more than 50 documents
- Volume processing incurs significant additional fees

KEY CATEGORY 5

Vendor Partnership & ROI (10%)

The question: Are you buying a roadmap or just a license?

Why It Matters

A legal AI tool's success depends heavily on the vendor's partnership mindset. You want a collaborator invested in your success, not just a software provider who disappears after the contract is signed. How the vendor behaves during the pilot is the best predictor of how they'll behave after.

What Good Looks Like

- Bespoke onboarding tailored to your team
- Rapid iteration on feedback during the pilot
- Transparent roadmap discussions
- Documented time savings with clear ROI calculation

Red Flags

- Vendor takes 48+ hours to respond during pilot
- One-size-fits-all demo with no practice-specific examples
- Hidden fees not disclosed until late in evaluation

Making the Decision

After the pilot, convert your findings into a crisp memo that you and your stakeholders can understand and trust.

Decision Framework

Recommendation	When to Use
Adopt	Weighted average score of 4.0+ across categories. Clear, documented ROI.
Proceed with Guardrails	Weighted average of 3.0-3.9. Value is present but some categories need oversight.
Do Not Adopt	Weighted average below 3.0, or accuracy scores 2 or below.

Your One-Page Decision Memo

1. **Tool evaluated:** Name, version, pilot dates
2. **Participants:** Number, practice groups, seniority mix
3. **Weighted scores**: By category and overall
4. **Top strengths:** 2-3 specific, documented wins
5. **Key limitations:** 2-3 concerns with mitigation plans
6. **ROI estimate:** Hours saved × blended rate
7. **Recommendation:** Adopt / Proceed with Guardrails / Do Not Adopt
8. **Proposed rollout:** Which practice groups, timeline, training plan

August v/s These Criteria

You've just read a framework for evaluating legal AI. Here's how we built August around those same criteria.

Accuracy & Reliability

Every August output includes citations to source documents, including audio files. When you ask a question about a contract, you can click through to the exact clause. When August extracts data across a hundred documents, each cell links back to the original text. We built a citation engine into the core architecture because lawyers need to verify before they trust.

August also knows when to stop. When the tool is uncertain or a question falls outside what the documents support, it says so rather than generating confident-sounding nonsense.

Firm Knowledge & Precedent Integration

Playbooks are central to how August works. You can build them from scratch, convert existing Word-based playbooks into structured rules, or extract them automatically from your precedent contracts. When you run a Playbook on a new agreement, August labels each clause as acceptable, needs review, or not acceptable, and proposes fallback language based on your standards.

This isn't generic AI reviewing contracts against generic standards. It's your institutional knowledge, captured once, applied consistently across every review.

We also integrate directly with SharePoint and document management systems. No manual uploads. No waiting for sync. August searches across your repositories in real time.

Platform Integration & Workflow Fit

We started with a different premise than most legal AI: the tool should live where lawyers already work.

The Word Add-in lets August edit documents even larger than 100 pages, working section by section with tracked changes. New content appears tracked so you can see exactly what changed. Your Playbooks apply directly inside Word.

August for Outlook turns your inbox into a workspace: summarize threads, review attachments without downloading, draft replies with full matter context. You can also search and pull emails directly from the main platform.

Context carries across all three environments. Start a task in Outlook, continue it in Word, reference it later in the web app. No re-uploading. No copy-pasting between windows.

Advanced Capabilities

Tabular Review extracts information across document sets into a sortable, filterable grid. You can chat directly with a review: "Which contracts have uncapped indemnities?" "Show all leases expiring in 2026." Full Text view shows entire clauses without clicking into each cell.

Workflow Mode lets you save any sequence of tasks and run them with one click. Upload a set of NDAs, extract key terms, flag deviations from your playbook, generate a summary memo. Save it once, run it forever.

Deep Research handles complex, multi-step investigations. Instead of running a single search, it iteratively identifies relevant documents, refines its approach, and expands coverage. All findings include citations to source material.

Vendor Partnership

We work as a service partner, not just a software vendor. That means bespoke onboarding tailored to your practice areas, rapid response when you have questions, and collaboration on building playbooks and workflows.

What This Looks Like in Practice

Task	Without August	With August
Summarize a complaint	1-3 hours	10-20 seconds
Identify obligations across 100+ contracts	10+ hours	~5 minutes
Analyze counterparty redlines	1-3 hours	30-50 seconds
Identify inconsistencies in witness testimony	15+ hours	30-45 seconds
Draft request for production	3-6 hours	30-50 seconds

To learn more or schedule a pilot, visit **www.august.law**