MPL LEGAL TECH ADVISORS

The 7-Step Legal AI Evaluation Framework

HOW TO MEASURE ACCURACY, EFFICIENCY AND COMPLIANCE

Why We Created This Framework

According to the **Legal Industry Report 2025 by FedBar**, only **21% of firms** have formally adopted generative AI, mostly in pilots or limited departments due to confidentiality and compliance concerns. [read here]

At the same time, **31% of individual lawyers** are already using AI weekly for drafting, research, and intake - often without approval. That gap is known as **Shadow AI**, and it puts client privilege, compliance, and malpractice coverage at risk.

AI adoption is already happening inside your practice. The only defensible step is to **evaluate and control it** before it creates liability.

The two pillars of safe legal AI are Compliance and Evaluation. See the break-downs of both below:

[Legal AI Evaluations Framework]

[Legal AI Compliance Framework]

Based on proven testing practices adapted for legal work, this framework shows you how to assess AI tools across three dimensions: **Technical Performance, Practice Value, and Risk & Compliance.** Stop gambling on AI vendor promises and start measuring what matters.

Book a 15min AI & Data Readiness Check

Part 1: The Three-Layer Evaluation Model

"You can't manage what you don't measure."

1.1. WHY TRADITIONAL AI EVALUATIONS FAIL IN LEGAL

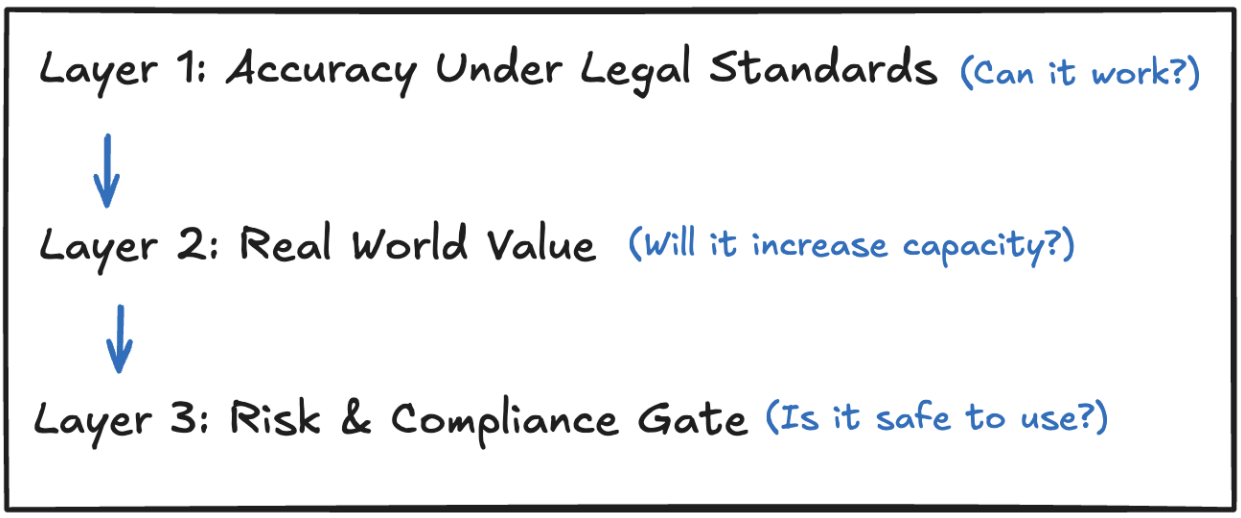
Most AI evaluations make one of three fatal errors:

1. **The demo trap:** evaluating based on AI vendor demonstrations rather than real firm data
2. **The accuracy illusion**: testing only technical performance without business context
3. **The compliance afterthought**: adding security checks after implementation instead of before

This framework addresses all three by creating a structured evaluation pipeline that mirrors how successful law firms actually adopt technology.

1.2. THE 3 LAYERS THAT MATTER

∗Each layer has specific pass/fail criteria. A tool must pass all three to be implementation-ready.



Book a 15min AI & Data Readiness Check

Part 2: Accuracy Under Legal Standards (Layer 1)

"If a legal tool can't prove where its answers come from, it doesn't matter how good the presentation looks."

2.1. WHY TRADITIONAL AI EVALUATIONS FAIL IN LEGAL

Core Principle: Your senior attorneys (not AI vendors or AI engineers) must validate the output quality.

STEP 1: CREATE YOUR TEST DATASET

Build an evaluation set using YOUR firm's actual work:

- 20-30 real documents from closed matters (redacted as needed)
- Include edge cases and complex scenarios you actually face
- Mix document types (contracts, briefs, research memos)

STEP 2: THE 3 CRITICAL TESTS

A. Cite or refuse (citation verification)

Ask the AI questions where answers exist in your documents

- **Pass:** Provides accurate quotes with source locations
- **Fail:** Gives answers without verifiable sources

B. Unanswerable drill (hallucination trap test)

Ask questions your documents cannot answer

- **Pass:** States "insufficient information" or similar
- **Fail:** Invents plausible-sounding information

C. Jurisdiction awareness

Provide mixed jurisdiction documents (e.g., EU and US law)

- **Pass:** Correctly identifies and separates jurisdictional differences
- **Fail:** Conflates different legal frameworks

STEP 3: SCORING FRAMEWORK

Score	Accuracy rate	Human review required	Status
Green	> 95% accurate	Spot checks only	Yes, with protocols
Yellow	85-94% accurate	All outputs reviewed	Pilot with oversight
Red	< 85% accurate	Not viable	No, reject or remediate

Key Insight: Even 90% accuracy means 1 in 10 outputs (or 100 out of 1000) contains errors, unacceptable for client work without review protocols.

Book a 15min AI & Data Readiness Check

Part 3: Real World Value (Layer 2)

"Technical accuracy means nothing if the tool doesn't actually save time, reduce cost or errors, or create capacity for client work."

This layer quantifies impact using measures that matter in legal context.

3.1. TIME TO VALUE CALCULATION

FORMULA

$$\text{Net Value} = (\text{Time Saved} \times \text{Hourly Rate} \times \text{Accuracy Rate}) - (\text{Review Time} + \text{Tool Cost})$$

EXAMPLE CALCULATION

- AI drafts contract clause in 5 minutes (vs. 30 minutes manual)
- Associate rate: $250/hour
- Time saved: 25 minutes = ~$105
- AI accuracy: 90% (requires 5-minute review)
- Review time cost: ~$20
- Net value per use: ~$73
- At 10 uses/week: ~$38,000 annual value

THE ADOPTION REALITY MATRIX

Metric	Green (Thriving)	Yellow (Struggling)	Red (Failing)
User adoption	> 75% regular use	40 - 74% use	< 40% use
Time to proficiency	< 1 week	1 - 4 weeks	> 1 month
Work interruptions	No interruptions	Minor adjustments	Major process changes
Technical issues	< 1/month	1 - 5/month	> 5/month

HIDDEN COST FACTORS:

1. **Wasted hours on workarounds**: the extra time staff spend fixing or re-doing tasks when an AI tool doesn't deliver.
2. **Interrupted focus**: lost efficiency when lawyers have to jump between multiple systems instead of working in one flow.
3. **Rebuilding confidence**: the time it takes your team to trust a tool again after it produces errors or misses key details.
4. **Ongoing retraining**: each update forces staff to relearn the system, costing time and slowing adoption.

Book a 15min AI & Data Readiness Check

Part 4: Risk & Compliance Gate (Layer 3)

“If you can’t defend the output to a regulator or insurer, you’re the one holding the liability.”

4.1. DATA HANDLING VERIFICATION PROTOCOL

THE FIVE ESSENTIAL QUESTIONS

**All must be yes to proceed*

1. **Jurisdiction lock:** is data processing location specified and fixed?
    a. **Safe:** *"Processed only in [specific country/region]"*
    b. **Unsafe:** *"Global cloud infrastructure"*
2. **Training exclusion:** written confirmation your data won't train models?
    a. **Safe:** *Explicit "No" in contract*
    b. **Unsafe:** *"May improve our services"*
3. **Deletion control:** can you delete data on demand?
    a. **Safe:** *Delete yourself via dashboard*
    b. **Unsafe:** *"Contact support for deletion requests"*
4. **Audit trail:** complete logs of access and processing?
    a. **Safe:** *Exportable logs with timestamps and action trace*
    b. **Unsafe:** *"Summary reports available"*
5. **Compliance attestation:** SOC 2 Type II or equivalent?
    a. **Safe:** *Current audit report provided*
    b. **Unsafe:** *"We follow best practices"*

4.2. PROFESSIONAL RESPONSIBILITY ALIGNMENT

Map to your jurisdiction's ethical rules:		
Rule category	**Requirement**	**Verification method**
Competence	Understand AI's limitations	Document training completed
Confidentiality	Protect client information	DPA signed & verified
Communication	Inform clients of AI work	Engagement letter updated
Supervision	Oversee AI assisted work	Review protocols documented

4.3. EMERGENCY SAFEGUARD

Every AI implementation needs:

- ➔ **An immediate shut-off option**: you have the power to stop the use at once if risks appear.
- ➔ **Data export rights:** the ability to retrieve all of your firm's information on demand.
- ➔ **Rollback procedures:** a clear way to undo changes or return to your prior system.
- ➔ **A ready backup process:** client work continues even if the tool is pulled.

Book a 15min AI & Data Readiness Check

Part 5: The Unified Evaluation Scorecard

"A proper evaluation ends with a defensible record and not just a gut feeling."

5.1. BRINGING IT ALL TOGETHER

Prepare a one-page assessment summary for each AI tool:

LEGAL AI TOOL EVALUATION SCORECARD

Tool: [*Name*] **| Use Case:** [*Primary purpose*] **| Date:** [*Evaluation date*]

TECHNICAL PERFORMANCE

- ➔ ***Accuracy score:*** _____% (*Target: > 95%*)
- ➔ **Hallucination test:** *Pass / Fail*
- ➔ **Citation capability:** *Yes / No*
- ➔ **Senior attorney approval:** *Yes / No*

Overall technical score: *GREEN / YELLOW / RED*

BUSINESS IMPACT

- ➔ **Time saved / week:** _____*hours*
- ➔ **Annual value estimate:** _____*$*
- ➔ **Adoption rate:** _____*%*
- ➔ **Workflow fit:** *LOW / MEDIUM / HIGH*

Overall impact score: *GREEN / YELLOW / RED*

RISK & COMPLIANCE

- ➔ **Data handling:** *Pass / Fail*
- ➔ **Ethics alignment:** *Compliant / Issues*
- ➔ **Audit trail:** *Complete / Partial / None*
- ➔ **Kill switch:** *Ready / Not ready*

Overall risk: *GREEN / YELLOW / RED*

FINAL RECOMMENDATION

- ☐ *PILOT*
- ☐ *REJECT*
- ☐ *REMEDIATE*

Notes: _____

Next review date: _____

Book a 15min AI & Data Readiness Check

Part 6: Practical Evaluation Steps

6.1. VALIDATE ACCURACY, VALUE, AND COMPLIANCE

Week 1: Proving accuracy

- Collect sample matters and documents for dataset creation
- Run accuracy checks on real legal tasks
- Senior attorney review of outputs

Week 2: Demonstrating value

- Measure time saved on routine tasks
- Calculate cost and capacity impact
- Gather attorney and staff feedback

Week 3: Compliance & risk review

- Review vendor security and data handling
- Map against professional ethics rules
- Conduct firm-wide risk assessment

Week 4: Final decision & record

- Prepare a one-page evaluation summary
- Present findings to partners
- Plan adoption or formally reject tool

6.2. ONGOING REVIEW & REFINE PROCESS

➔ **Quarterly reviews:** *re-check accuracy using fresh matters and documents*

➔ **Monthly metrics:** *compare actual time and cost savings against projections*

➔ **Incident tracking:** *record and review every error or issue for compliance purposes*

➔ **Feedback integration:** *incorporate user input to refine use*

Book a 15min AI & Data Readiness Check

Part 7: Red Flags

7.1. TECHNICAL RED FLAGS

- ➔ No ability to show sources or citations
- ➔ Accuracy below 85% when tested on your own matters
- ➔ Results vary widely across practice areas
- ➔ Cannot restrict outputs to your firm's templates

7.2. BUSINESS RED FLAGS

- ➔ Setup or rollout takes longer than 3 months
- ➔ Demands major changes to existing workflows
- ➔ Fewer than 50% of users adopt the tool after training
- ➔ Hidden costs exceed 30% of the quoted price

7.3. COMPLIANCE RED FLAGS

- ➔ Vendor refuses to complete a security questionnaire
- ➔ No option to block use of your data for model training
- ➔ Data stored or processed outside your chosen jurisdiction
- ➔ Vendor lacks professional liability insurance coverage

Book a 15min AI & Data Readiness Check

Part 8: Legal AI Vendor Evaluation Conversation

8.1. QUESTIONS THAT REVEAL THE TRUTH

- ➔ For **accuracy claims**: “Can you run our test dataset through your system right now and show us the outputs?”
- ➔ For **integration promises**: “Can you specify another firm using your tool with our practice management system?”
- ➔ For **security assurances**: “Will you indemnify us for any data breach or confidentiality
- ➔ Violation?”
- ➔ For **evidence of impact:** “Can we speak to three references who’ve achieved these results?”

8.2. WARNING SIGNS IN VENDOR RESPONSES

- ➔ "Our proprietary technology..." = **avoiding specifics**
- ➔ "Most firms see..." = **no concrete examples**
- ➔ "After full installation..." = **moving goalposts**
- ➔ "Our AI is different..." = **deflecting standard requirements**

Book a 15min AI & Data Readiness Check

Quick Reference: The 7-Step Summary

1. **Map your test set:** Use 20–30 real (redacted) matters and documents, including edge cases.
2. **Prove accuracy:** Require citations, test unanswerable questions, and check jurisdiction handling.
3. **Quantify value:** Measure time saved, calculate cost/capacity impact, and gather attorney feedback.
4. **Check compliance:** Demand written answers on jurisdiction, training exclusions, deletion rights, and audit logs.
5. **Document findings:** Create a one-page scorecard for every tool, covering accuracy, impact, and risk.
6. **Decide with evidence:** Pilot, remediate, or reject based on the full record, not vendor demos.
7. **Review regularly:** Re-check quarterly with new data, track incidents, and update as rules evolve.

Conclusion: From Evaluation to Excellence

The difference between law firms and in-house teams that successfully adopt AI and those that fail is about evaluation rigor. This framework transforms AI adoption from a leap of faith into a measured, strategic decision.

Remember:

1. Bad evaluation = Bad outcomes
2. You can't manage what you don't measure
3. Every "NO" based on data protects your privilege
4. Every "YES" based on evidence grows your practice

Your next actions:

1. **Immediate:** Inventory current AI tools using this framework
2. **This week:** Create your test dataset from real matters
3. **This month:** Evaluate one high-priority AI tool completely
4. **This quarter:** Establish regular evaluation cycles

The goal isn't to avoid AI. It's to use it without becoming the next cautionary tale.

This blueprint implements the 7-step legal AI evaluation framework, incorporating lessons from proven LLMOps best practices and extensive work with law firms & in-house teams. For guidance specific to your firm, visit [our website](#).