

LEGAL AI TOOL

EVALUATION SCORECARD

Compare Westlaw, Lexis+ AI, vLex and Custom-Built Assistants

A weighted, vendor-neutral framework for choosing the right AI legal research tool

Inside this scorecard

- **A requirements profile worksheet and a build-vs-buy decision guide**
- **7 weighted evaluation categories with criteria, vendor questions and red flags**
- **A fillable master scorecard and a hands-on citation test protocol**
- **A 25-question vendor demo bank and a recommendation memo template**

Built for Forward Deployed Engineers and legal technology teams

How to Use This Scorecard

Choosing an AI legal research tool is a high-stakes decision. A tool that writes fluent answers but invents a citation can create professional risk for every lawyer who relies on it. This scorecard gives your evaluation panel a shared, structured way to compare options on the factors that actually matter in legal work: source coverage, citation reliability, security, workflow fit, adoption, governance and cost.

The framework is vendor-neutral. It does not pre-score any product, because features, content coverage and AI capabilities vary by subscription, jurisdiction and release. Your team scores each option based on its own demos, trials and tests.

The 5-Step Evaluation Process

Step	What you do	Output	Typical time
1. Define needs	Complete the Requirements Profile Worksheet with lawyers, KM and IT.	Signed-off requirements and must-haves	1 week
2. Shortlist	Use the Build vs Buy guide to select 2 to 4 options.	Shortlist with rationale	1 week
3. Hands-on test	Run the Test Protocol on each option using real research questions.	Completed Citation Verification Log	2 to 3 weeks
4. Score	Each panel member scores independently, then the panel agrees final scores.	Completed Master Scorecard	1 week
5. Decide	Write the Recommendation Memo and define the pilot.	Approved decision and pilot plan	1 week

The Scoring Scale (1 to 5)

Score every criterion on the same five-point scale. Write a short evidence note next to each score, such as "3 of 20 test citations could not be verified", so the panel can defend its decision later.

Score	Label	Meaning
5	Excellent	Fully meets the requirement, proven in your own testing, no material gaps.
4	Strong	Meets the requirement with minor gaps that have a clear workaround.
3	Adequate	Partially meets the requirement; gaps need configuration, process changes or extra cost.
2	Weak	Significant gaps that would affect daily use or create risk.
1	Poor / Absent	Does not meet the requirement, or the vendor could not demonstrate it.

How Weighting Works

Not every category matters equally. Each of the 7 categories carries a weight, and all weights add up to 100%. A category score is the average of its criterion scores. The weighted score is the category score multiplied by its weight.

Worked example

Tool A averages 4.2 on Citation Reliability (weight 20%) and 3.0 on Cost (weight 10%).

Citation Reliability: $4.2 \times 0.20 = 0.84$

Cost: $3.0 \times 0.10 = 0.30$

Add the weighted scores for all 7 categories to get a total out of 5.00. A total of 4.0 or higher indicates a strong fit; below 3.0 usually means the option should not proceed to pilot.

The default weights reflect a typical law firm. Adjust them to your context using the "Your weight" column in the Master Scorecard. For example, a firm handling highly confidential M&A work may raise Security to 25% and reduce Usability to 5%.

Who Should Sit on the Evaluation Panel

A legal AI decision made only by IT, or only by partners, usually fails at adoption or at security review. Build a small cross-functional panel of 4 to 6 people:

- Practising lawyer (senior): judges legal accuracy, relevance and whether the reasoning holds up.
- Research associate or paralegal: the heaviest daily user; judges speed, usability and fit with real research tasks.
- Knowledge management lead: judges how well the tool surfaces internal know-how, precedents and templates.
- IT / information security: judges integration, access control, data handling and audit capability.
- Procurement or finance: judges pricing models, contract terms and total cost of ownership.
- Forward Deployed Engineer (if available): translates requirements into technical tests and evaluates build options.

Step 1: Requirements Profile Worksheet

Before looking at any product, document what your organisation actually needs.

Vendors run polished demos; a written requirements profile keeps the panel focused on your workflows rather than the vendor's best features. Complete this worksheet in a workshop with the panel and get it signed off before Step 2.

Requirement	Guidance and example	Your answer
Jurisdictions	List every jurisdiction you research regularly. Example: "Primary: England & Wales. Secondary: EU, Singapore."	
Practice areas	List priority practice areas for the first rollout. Example: "Commercial contracts, regulatory compliance, employment."	
Number of users	Split by role. Example: "40 lawyers, 12 paralegals, 5 KM staff."	
Internal document volume	Estimate internal opinions, memos and precedents to search. Example: "About 18,000 research memos, 2,500 precedent clauses."	
Data sensitivity level	Classify data the tool will touch. Example: "Client-privileged; some matters under strict ethical walls."	
Existing systems	Document management, intranet, productivity suite, matter management. Example: "iManage, Microsoft 365, SharePoint intranet."	

Current research tools	What subscriptions exist today and when they renew. Example: "Licensed database, renews March."	
Budget range	Annual budget including implementation. Example: "Year 1: set range; ongoing: per-seat ceiling."	
Go-live timeline	Target pilot and rollout dates. Example: "Pilot in Q1, firm-wide by Q3."	
Data residency	Where data must be stored and processed. Example: "Must remain in-region."	

Must-Have vs Nice-to-Have Checklist

Mark each capability as a Must-have (M) or Nice-to-have (N). Any option that fails a Must-have is removed from the shortlist, regardless of its total score. Keep Must-haves to a short list; if everything is a Must-have, nothing is.

Capability	M / N	Example of why it matters
Pinpoint citations to page or paragraph	☐ M ☐ N	Lawyers need to verify the exact passage, not just the case name.

Flags overruled or outdated authorities	☐ M ☐ N	Relying on overturned law is a serious professional risk.
Searches internal documents alongside licensed content	☐ M ☐ N	Reusing past research avoids duplicate work across teams.
Respects ethical walls and matter-level permissions	☐ M ☐ N	A conflict-walled lawyer must never see restricted matter files.
Single sign-on and role-based access	☐ M ☐ N	Simplifies access management and offboarding.
Full audit log of queries and outputs	☐ M ☐ N	Supports supervision, quality review and regulatory enquiries.
Jurisdiction filter on every query	☐ M ☐ N	Prevents mixing authorities from the wrong jurisdiction.
Export to research memo template	☐ M ☐ N	Saves time when converting findings into client work.
Integration with document management system	☐ M ☐ N	Keeps research inside existing matter workspaces.

Contractual commitment not to train on firm data	☐ M ☐ N	Protects client confidentiality and privilege.
--	---	--

Step 2: Build vs Buy Decision Guide

There are three realistic paths. Commercial platforms bring curated legal content and editorial features. Custom builds bring control over internal knowledge and workflows. Hybrid approaches combine both. Use the questions below to decide which paths belong on your shortlist.

Decision Questions

Answer each question in order. Your answers point to the most suitable path.

1. Are licensed legal databases sufficient for most research? If yes, a commercial platform is the natural starting point. If your lawyers rely heavily on internal opinions, past advice and firm precedents, continue to question 2.
2. Do you need to search internal opinions alongside licensed content in one interface? If yes, consider a hybrid or custom path. Many firms find their most valuable knowledge sits in their own document management system.

3. Do you have strict data-residency, privilege or client-mandated hosting rules? If yes, confirm whether commercial platforms meet them. If they do not, a custom build in your own environment may be required.
4. Do you have engineering capacity to build and maintain an AI system? A custom RAG assistant needs ongoing work on retrieval quality, evaluation, security and model updates. Without an FDE or equivalent team, a custom build carries delivery risk.
5. Can licensed content be used inside a custom system? Check database licence terms carefully. Many licences restrict exporting or indexing content into third-party AI systems.

Comparison of the Three Paths

Factor	Commercial platform	Hybrid (platform + custom layer)	Fully custom RAG assistant
Legal content	Curated, editorially maintained content included	Platform content plus internal documents	Only content you are licensed and able to index

Internal knowledge	Limited or add-on dependent	Strong: custom layer searches internal sources	Strong: designed around your sources
Time to value	Fast: weeks	Medium: 2 to 4 months	Slower: 3 to 6+ months
Control over behaviour	Low to medium	Medium	High
Engineering effort	Low	Medium	High, ongoing
Data control	Vendor-managed under contract	Shared	Full, within your environment
Cost profile	Subscription per seat	Subscription plus build and run costs	Build, hosting, model usage and maintenance

When Each Option Makes Sense

Commercial platform: example

A 30-lawyer firm focused on domestic commercial litigation, with no in-house engineers and good existing database subscriptions. Its research is mostly

case law and statute. A commercial platform delivers value quickly with the least risk.

Hybrid: example

A 200-lawyer firm with 15 years of internal research memos in its document management system. Lawyers want licensed case law and the firm's own past advice in one answer. The firm keeps its platform subscription and an FDE builds a secure internal-knowledge layer that links to platform results.

Fully custom: example

An in-house legal department at a regulated bank, researching mainly its own policies, regulator guidance and prior legal opinions, with strict rules that data must stay inside the bank's cloud tenancy. A custom RAG assistant, built and evaluated by an FDE team, fits these constraints best.

Step 3: The 7 Evaluation Categories

Each category below includes the criteria to score, a practical example, questions to ask, red flags and a description of what good looks like. Score every criterion from 1 to 5 using the scale on page 2, and record your evidence in the Master Scorecard.

#	Category	Default weight
1	Source Coverage & Jurisdiction Support	20%
2	Citation Reliability & Answer Accuracy	20%
3	Security, Confidentiality & Access Control	15%
4	Integration & Workflow Fit	15%
5	Usability & Lawyer Adoption	10%
6	Governance, Licensing & Vendor Terms	10%
7	Cost & Total Cost of Ownership	10%
	Total	100%

Category 1: Source Coverage & Jurisdiction Support

(Default weight: 20%)

An AI assistant can only be as good as the sources behind it. This category checks whether the tool has the primary and secondary law your lawyers need, for the jurisdictions they work in, and whether that content is current.

Criteria to score

#	Criterion	What you are checking
1.1	Primary law coverage	Case law, legislation and regulations for your priority jurisdictions and courts.
1.2	Secondary sources	Commentary, practice guides, treatises and journals relevant to your practice areas.
1.3	Currency of content	How quickly new judgments and legislative changes appear, and whether amended versions are tracked.
1.4	Jurisdiction filtering	Ability to restrict every query to a chosen jurisdiction and court level.
1.5	Internal content support	Ability to search approved internal memos, opinions and precedents.

1.6	Metadata quality	Accurate tagging of court, date, case name, topic and status.
-----	------------------	---

Example

A firm tests three tools with the question: "What remedies are available for repudiatory breach of a supply agreement?" Tool A returns strong domestic authorities. Tool B mixes in authorities from another jurisdiction without labelling them. Tool C cannot search the firm's own past advice on the same topic. Tool A scores 5 on jurisdiction filtering; Tool B scores 2.

Questions to ask the vendor (or your build team)

- Which courts, tribunals and regulators are covered for our jurisdictions, and from what date?
- How long, on average, between a judgment being published and it appearing in the tool?
- Does the AI feature search the same content as the core platform, or a subset?
- Can we restrict answers to a single jurisdiction, and is that restriction enforced or only suggested?

- Can the tool index our internal documents, and how are they kept in sync?

Red flags

- The AI feature only covers a subset of the platform's content, and this is not clearly disclosed.
- No clear labelling of which jurisdiction an authority comes from.
- Vague answers about update frequency, such as "regularly updated".

What good looks like

Comprehensive, current coverage of your priority jurisdictions, an enforced jurisdiction filter on every query, clearly labelled sources, and the option to include approved internal documents.

Category 2: Citation Reliability & Answer Accuracy

(Default weight: 20%)

This is the category where legal AI tools succeed or fail. Every important claim must trace back to a real, relevant and current authority. A tool that produces confident but unsupported answers is a liability, regardless of how fast it is.

Criteria to score

#	Criterion	What you are checking
2.1	Citation existence	Every cited authority actually exists. Zero tolerance for fabricated cases.
2.2	Pinpoint accuracy	Citations link to the specific page, paragraph or section that supports the claim.
2.3	Quotation vs summary	The tool clearly separates direct quotations from its own summaries.
2.4	Status awareness	The tool flags overruled, reversed, superseded or repealed authorities.
2.5	Contrary authorities	The tool surfaces conflicting judgments rather than only supporting ones.
2.6	Honest uncertainty	The tool says when evidence is insufficient instead of guessing.

Example

In testing, a lawyer asks about the enforceability of a limitation clause. Tool A cites five cases with links to the exact paragraphs, and notes that one appellate decision takes a narrower view. Tool B cites six cases, but one cannot be found in any database and another was overturned on appeal without any warning. Tool A scores 5 on citation existence and contrary authorities; Tool B scores 1 on citation existence.

Questions to ask the vendor (or your build team)

- How does the system verify that a citation exists before showing it?
- Does every claim link to a source passage, or only the answer as a whole?
- How does the tool show whether an authority is still good law?
- What happens when the tool cannot find enough support? Can you show us an example?
- Do you publish accuracy or evaluation results, and how were they measured?

Red flags

- Any fabricated citation during your testing.
- Answers that cite sources only at the end, with no link between specific claims and specific authorities.
- The tool never says "I could not find support for this".
- Status warnings that only exist in the core platform, not in AI-generated answers.

What good looks like

Every claim is linked to a verifiable passage; quotations are marked; overruled authorities are flagged; contrary views are surfaced; and the tool openly states when the evidence does not support a conclusion.

Category 3: Security, Confidentiality & Access Control

(Default weight: 15%)

Legal research queries often reveal client strategy, deal details or litigation positions.

The tool must protect privileged information, respect internal permissions and give the firm visibility into how data is used.

Criteria to score

#	Criterion	What you are checking
3.1	Data use commitments	Contractual commitment that firm data and queries are not used to train shared models.
3.2	Encryption	Data encrypted in transit and at rest, with clear key management.
3.3	Access control	Single sign-on, role-based access and enforcement of ethical walls.
3.4	Audit logging	Complete logs of users, queries, sources retrieved and outputs.
3.5	Data residency and retention	Control over where data is stored and how long it is kept.
3.6	Prompt-injection safeguards	Protection against malicious instructions hidden inside retrieved documents.

Example

During the security review, a firm uploads a test internal memo to a restricted matter workspace. A lawyer outside that matter then asks a related question. Tool A excludes the memo entirely. Tool B includes a summary of the memo in its answer. Tool B fails the ethical-wall test and scores 1 on access control, which also fails a Must-have.

Questions to ask the vendor (or your build team)

- Are our queries, uploaded documents or outputs used to train or improve models for other customers?
- Where is data stored and processed, and can we choose the region?
- How are document-level and matter-level permissions enforced in AI answers?
- What is logged, who can access the logs, and how long are they kept?
- Which independent security certifications or audits do you hold, and can we see the reports?
- How do you defend against instructions embedded in documents the system retrieves?

Red flags

- **Data-use terms are only in a policy page, not in the contract.**
- **AI answers can draw on documents the user is not permitted to open.**
- **No audit log of AI queries and outputs.**

What good looks like

Clear contractual data commitments, encryption, single sign-on, enforced matter-level permissions, full audit logs, regional data control and tested defences against prompt injection.

Category 4: Integration & Workflow Fit (Default weight: 15%)

A tool that lives outside lawyers' daily workflow will be used less. This category checks how well the tool connects to existing systems and how naturally it fits the way your teams already research, draft and file work.

Criteria to score

#	Criterion	What you are checking
4.1	Document management integration	Connects to your DMS for search, saving results and respecting permissions.
4.2	Productivity suite integration	Works within word processing, email or collaboration tools lawyers already use.
4.3	APIs and extensibility	Documented APIs for custom workflows and internal tools.
4.4	Export options	Exports findings into research memo or brief templates with citations intact.
4.5	Research history and sharing	Saves research sessions and lets colleagues reuse them.

Example

A research associate completes a contract-termination research task. With Tool A, she saves the result directly to the matter workspace and exports a memo draft with citations. With Tool B, she copies text into a document manually and re-types citations. Tool A scores 5 on export; Tool B scores 2.

Questions to ask the vendor (or your build team)

- Which document management and productivity systems do you integrate with today, not on the roadmap?
- Do you provide APIs, and what can they access?
- Can research results be exported with citations and links preserved?
- Can we share a research session with a colleague or supervising partner?

Red flags

- Integrations listed as "coming soon" are counted as available.
- No API, or an API without legal content access.
- Exports that lose citations or formatting.

What good looks like

Seamless connection to your DMS and productivity tools, a usable API, one-click export into firm templates, and shareable research history.

Category 5: Usability & Lawyer Adoption (Default weight: 10%)

Even an accurate tool fails if lawyers do not use it. This category measures how easy the tool is to learn, how quickly it returns results and how confident lawyers feel using it in real work.

Criteria to score

#	Criterion	What you are checking
5.1	Ease of learning	Time for a new user to run a productive research query.
5.2	Interface clarity	Clear layout of answers, sources, filters and follow-up options.
5.3	Response speed	Time to return a useful, cited answer.
5.4	Follow-up and comparison	Ability to ask follow-ups, compare judgments and refine questions.
5.5	Training and support	Quality of onboarding, documentation and vendor support.

Example

Five associates each receive a 30-minute introduction to two tools and then complete the same three research tasks. Four of five complete all tasks in Tool A within 20 minutes; only two manage it in Tool B. Average satisfaction is 4.4 for Tool A and 3.1 for Tool B.

Questions to ask the vendor (or your build team)

- **What training do you provide during rollout and afterwards?**
- **Is there in-product guidance on writing effective research questions?**
- **What is the typical response time for a complex research question?**
- **What adoption rates do comparable firms see after six months?**

Red flags

- **Only senior or highly technical users can get good results.**
- **Support is limited to email with slow response times.**
- **The demo uses prepared questions that your team cannot reproduce.**

What good looks like

New users produce useful, cited research within their first session, the interface makes sources easy to inspect, and adoption stays strong after the pilot.

Category 6: Governance, Licensing & Vendor Terms

(Default weight: 10%)

Legal AI introduces contractual and governance questions that traditional research tools did not. This category checks whether the vendor relationship supports responsible, long-term use.

Criteria to score

#	Criterion	What you are checking
6.1	Content licensing clarity	Clear terms on how licensed content may be used, stored and shared.

6.2	Model change transparency	Notice when underlying models or retrieval methods change.
6.3	Liability and indemnity	Reasonable allocation of risk for errors and IP claims.
6.4	Exit and data portability	Ability to export your data and research history if you leave.
6.5	Vendor stability and roadmap	Financial stability, product commitment and credible roadmap.

Example

Two vendors offer similar features. Vendor A commits to 30 days' notice before changing the underlying model and lets the firm export all saved research. Vendor B can change models without notice and keeps research history on termination. The firm scores Vendor A 5 and Vendor B 2 on model change transparency and exit.

Questions to ask the vendor (or your build team)

- Will you notify us before changing the AI model or retrieval approach, and how much notice will you give?

- **What happens to our data, saved research and logs if we end the contract?**
- **Can licensed content be used inside our own internal tools or custom AI layer?**
- **What liability do you accept for incorrect AI outputs?**

Red flags

- **Silent model updates with no change log.**
- **No way to export saved research.**
- **Licence terms that prohibit linking platform content to internal tools you plan to build.**

What good looks like

Transparent change management, clear licensing for your intended use, fair risk allocation and a clean exit path with full data portability.

Category 7: Cost & Total Cost of Ownership (Default weight: 10%)

The subscription price is only part of the cost. Total cost of ownership includes implementation, integration, training, internal time, usage-based charges and ongoing maintenance, especially for custom builds.

Criteria to score

#	Criterion	What you are checking
7.1	Pricing model clarity	Clear per-seat, usage-based or enterprise pricing with no hidden charges.
7.2	Implementation cost	Setup, integration, data preparation and configuration effort.
7.3	Ongoing run cost	Hosting, model usage, support and maintenance over three years.
7.4	Internal effort	Time required from IT, KM, lawyers and engineers.
7.5	Value vs cost	Expected time savings and quality gains relative to total cost.

Example

Tool A costs more per seat but includes integration and training. A custom build has lower licence costs but needs an engineer for six months plus ongoing hosting and model usage. When the firm models three-year TCO, the gap between the options narrows from 40% to 12%, and the decision shifts to the value each provides.

Questions to ask the vendor (or your build team)

- **What is the full three-year cost, including implementation, training and expected price increases?**
- **Are there usage limits, and what happens when we exceed them?**
- **Is the AI feature included in our current subscription or charged separately?**
- **What internal staff time do comparable customers typically spend on rollout?**

Red flags

- Pricing that only covers year one.
- Usage caps that are not disclosed until contract stage.
- Custom build estimates that exclude evaluation, security work and maintenance.

What good looks like

A transparent three-year cost model covering all direct and internal costs, with value measured against a clear baseline of current research time.

Step 4: Master Scorecard

Record the agreed category score (average of criterion scores, 1 to 5) for each option, then multiply by the weight to get the weighted score. Use "Your weight" if you have adjusted the defaults; your weights must still total 100%. First confirm that every option passed all Must-haves from Step 1.

Must-Have Gate

Option	All Must-haves met?	Failed Must-haves (if any)
Westlaw	☐ Yes ☐ No	
Lexis+ AI	☐ Yes ☐ No	
vLex	☐ Yes ☐ No	
Custom build	☐ Yes ☐ No	

Category Scores

Enter scores as "Score / Weighted", for example "4.2 / 0.84".

#	Category	Default weight	Your weight	Westlaw	Lexis+ AI	vLex	Custom build
1	Source Coverage & Jurisdiction	20%					
2	Citation Reliability & Accuracy	20%					

3	Security & Access Control	15%					
4	Integration & Workflow Fit	15%					
5	Usability & Adoption	10%					
6	Governance & Vendor Terms	10%					
7	Cost & TCO	10%					
	Weighted total (out of 5.00)	100%	100%				

Evidence Notes

For each category, record the evidence behind the scores. Example: "Category 2, Tool B: 2 of 20 citations not verifiable; 1 overruled case with no warning."

Category	Key evidence and notes
----------	------------------------

Source Coverage & Jurisdiction	
Citation Reliability & Accuracy	
Security & Access Control	
Integration & Workflow Fit	
Usability & Adoption	
Governance & Vendor Terms	
Cost & TCO	

--	--

Evaluator	Role	Date	Initials

Interpreting the total

- **4.0 to 5.0: Strong fit. Proceed to pilot.**
- **3.0 to 3.9: Conditional fit. Proceed only if gaps have an agreed fix or workaround.**
- **Below 3.0: Weak fit. Do not pilot without major changes.**

A close total between two options (within 0.2) should be decided by Category 2 (Citation Reliability) and Category 3 (Security).

Hands-On Test Protocol

Demos show what a vendor wants you to see. Hands-on testing shows how a tool behaves on your work. Run the same set of questions through every shortlisted option, using real (anonymised) research questions from recent matters. Aim for at least 20 questions across the six types below.

Six Query Types to Test

Query type	Purpose	Example question
Straightforward	Checks basic retrieval and citation accuracy.	"What are the statutory requirements for a valid company resolution in [jurisdiction]?"
Complex	Checks multi-issue reasoning across several authorities.	"Can a supplier terminate for material breach where the contract has both a cure period and an exclusive remedies clause?"

Ambiguous	Checks whether the tool asks for clarification or states its assumptions.	"Is the clause enforceable?" (with no clause or jurisdiction given)
Conflicting judgments	Checks whether contrary authorities are surfaced.	"How do courts interpret 'reasonable endeavours' in supply contracts?"
Outdated law	Checks status awareness for repealed or overruled authorities.	A question that relies on a provision you know has been amended or a case known to be overruled.
Unanswerable	Checks honest uncertainty instead of fabrication.	A question about a fictitious statute or a very niche issue with no reported authority.

Tip

Include at least two "trap" questions, such as a fictitious case name. A reliable tool should say it cannot find the authority. Any tool that confidently summarises a case that does not exist should be scored 1 on citation existence.

Citation Verification Log

For every citation in every answer, a lawyer or research associate checks the original source and records the result. A citation passes only if it exists, supports the claim at the cited location and is still good law.

Query #	Tool	Citation given	Exists?	Correct passage?	Still good law?	Pass/Fail
Example: Q3	Tool A	Case X v Y, para 42	Yes	Yes	Yes	Pass
Example: Q3	Tool B	Case P v Q	No	n/a	n/a	Fail

Calculating Accuracy

- Citation accuracy % = $(\text{Citations passed} \div \text{Total citations checked}) \times 100$
- Hallucination count = Number of citations that do not exist
- Unsupported answer rate % = $(\text{Answers with one or more unsupported claims} \div \text{Total answers}) \times 100$
- Honest uncertainty rate % = $(\text{Unanswerable questions correctly declined} \div \text{Unanswerable questions asked}) \times 100$

Example result

Tool A: 94 of 100 citations passed (94% accuracy), 0 hallucinations, 2 of 2 trap questions correctly declined. Tool B: 81 of 100 passed (81%), 3 hallucinations, 0 of 2 trap questions declined. Tool B's hallucinations alone justify a score of 1 on Criterion 2.1.

Vendor Demo Question Bank

Send these 25 questions to vendors before the demo and ask for written answers.

Then ask them to prove the key answers live, using your own questions. Written

answers become useful evidence for scoring and for contract negotiation. For a custom build, put the same questions to your engineering team.

Source Coverage

- 1. Which jurisdictions, courts and regulators are covered, and from which dates?**
- 2. How quickly are new judgments and legislative changes added?**
- 3. Does the AI search all platform content or only part of it?**
- 4. Can we add our own internal documents, and how are they kept up to date?**

Citation Reliability

- 5. How do you prevent fabricated citations from appearing in answers?**
- 6. Can you show an answer where the tool says it could not find support?**
- 7. How are overruled or superseded authorities flagged in AI answers?**
- 8. Is every claim linked to a specific source passage?**

Security & Confidentiality

- 9. Are our queries or documents used to train models for other customers?**
- 10. Where is our data stored and processed, and can we choose the region?**
- 11. How are ethical walls and document permissions enforced in AI answers?**
- 12. What do your audit logs capture, and how long are they retained?**

13. How do you defend against prompt injection in retrieved documents?

Integration & Workflow

14. Which DMS and productivity integrations are live today?

15. What can your API access, and are there usage limits?

16. Can research be exported to our templates with citations intact?

Usability & Adoption

17. What onboarding and ongoing training do you provide?

18. What adoption rates do similar firms reach after six months?

19. Can we run a trial with our own questions and users?

Governance & Vendor Terms

20. How much notice do you give before changing the underlying model?

21. What happens to our data and research history if we leave?

22. What liability do you accept for incorrect outputs?

Cost & TCO

23. What is the full three-year cost including implementation and training?

24. Is the AI capability included or charged separately?

25. What usage limits or overage charges apply?

How to use the answers

Mark each answer as **Demonstrated (shown live)**, **Stated (written claim only)** or **Unclear**. Give more weight to demonstrated capabilities when scoring. A stated capability that cannot be demonstrated during the trial should be scored no higher than 3.

Step 5: Decision Summary & Recommendation Memo

Turn your scores and evidence into a short memo for decision-makers. Keep it to two pages. Partners and general counsel need the recommendation, the evidence and the risks, not every individual score. Use the template below, with the example answers as a guide.

Section	What to include	Example
1. Purpose	The decision being requested.	Approve a 3-month pilot of Tool A for the commercial and regulatory teams.

2. Requirements summary	Top 5 requirements and Must-haves from Step 1.	Domestic case law, internal memo search, ethical walls, audit logs, DMS integration.
3. Options considered	Shortlisted options and why others were excluded.	Tools A, B and C plus a hybrid option. Tool C excluded: no regional data hosting.
4. Scores	Weighted totals and the deciding categories.	Tool A 4.3, hybrid 4.0, Tool B 3.2. Tool A leads on citation reliability (4.8).
5. Key test results	Citation accuracy, hallucinations, trap questions.	Tool A: 94% citation accuracy, 0 hallucinations, 2 of 2 trap questions declined.
6. Risks and mitigations	Main risks and how they will be managed.	Limited internal-document search; mitigated by a hybrid layer in phase 2.
7. Recommendation	The recommended option and rationale.	Proceed with Tool A; evaluate hybrid extension after the pilot.
8. Pilot scope	Teams, users, duration, budget.	25 users across two teams for 12 weeks.
9. Success metrics	How success will be measured.	See the pilot metrics table below.
10. Sign-offs	Approvers and dates.	Managing partner, head of KM, CISO, finance director.

Pilot Success Metrics

Measure a baseline before the pilot starts, then compare at the end. Decisions to scale should rest on measured results, not impressions.

Metric	How to measure	Example target
Research time saved	Average time to complete standard research tasks, before vs during pilot.	30% reduction
Citations verified	Percentage of AI-provided citations that pass the verification log.	95% or higher
Unsupported answer rate	Percentage of answers with any unsupported claim.	Below 5%
Lawyer satisfaction	Survey score on usefulness and trust (1 to 5).	4.0 or higher
Active usage	Percentage of pilot users using the tool weekly.	70% or higher

Remember

AI tools support legal research; they do not replace professional judgment.

Whichever option you choose, lawyers remain responsible for verifying authorities and approving any advice or filing based on AI-assisted research.

Re-run parts of this evaluation whenever the vendor changes models, content sources or access permissions.

GSDC CERTIFIED FORWARD DEPLOYED ENGINEER

Evaluating legal AI is an FDE skill.

Building it is the next step.

The scorecard you just used reflects how a Forward Deployed Engineer works: understand the users, define measurable requirements, test rigorously and deploy safely. The GSDC Forward Deployed Engineer Certification builds the skills to take the next step: designing and delivering AI solutions like a legal research assistant, end to end.

What You Will Learn

Skill area	How it applies to legal AI
Retrieval-augmented generation (RAG)	Ground answers in approved legal sources with hybrid search and reranking.
LLM evaluation	Build test sets, measure citation accuracy and catch hallucinations before go-live.
Systems integration	Connect AI to document management, databases and productivity tools via APIs.
Secure deployment	Apply access controls, audit logging and prompt-injection safeguards.
Stakeholder discovery	Translate lawyers' workflows into measurable technical requirements.

CERTIFIED FORWARD DEPLOYED ENGINEER

LEARN DIRECTLY FROM INDUSTRY EXPERTS, AI PRACTITIONERS, AND ENGINEERING LEADERS WHO ARE BUILDING AND DEPLOYING SCALABLE AI DRIVEN SOLUTIONS FOR ENTERPRISES WORLDWIDE.



ABOUT GSDC CERTIFICATION



EBOOK

Extensive and exclusive Ebook created by world's experts to help you with understanding core concepts.



EBOOK

Extensive and exclusive Ebook created by world's experts to help you with understanding core concepts.



CREATED BY EXPERTS

GSDC certifications are created and authored by world's leading experts in the field.



LEARNING MATERIALS

Get access to learning materials such as videos, ebooks, templates, and practice exams, which will help you clear the certification exam.

LEARNING OBJECTIVE

- Validate your Forward Deployed Engineering expertise.
- Increase your earning potential with an in-demand credential.
- Get complimentary career assistance after certification.

Enroll now with the code **LEARN20** To avail **20%** discount

Enroll Now



www.gsdccouncil.org