AI Governance Guide

For Law Firms and Legal Departments

What governance actually requires once the policy is written

By Colin S. Levy

Foreword

This primer answers two questions for working lawyers. What does an AI governance program look like for a law firm or legal department, and where do you start? And once the policy is signed, the committee is chartered, and the approved tool list is published, why do so many programs still struggle?

The first question is the easier one. The professional-responsibility floor is reasonably well-mapped, the NIST framework gives the architecture, and the major regulatory regimes have settled enough that a defensible starting position is within reach for any size practice. Part I sets that out.

The second question is harder, and Part II is where most of this primer's utility lives. AI governance is only fractionally a documents problem. The other fraction, which is most of it, is about behavior, judgment, incentives, model literacy, and the daily friction of asking lawyers to add a verification step to work that already takes too long. Why lawyers route around rules they helped write, why verification controls degrade within a quarter, why two practices with identical policies produce wildly different outcomes: those are the chapters that follow.

Read Part I to know what to do. Read Part II to know why most programs still underperform once they have done it.

Contents

Part I sets out the foundational material a working program needs. Part II is where most of the added utility lives, addressing the operational questions that surface once the policy is in place.

Part I

Foundations

The chapters in Part I are the foundation a working program needs: the regulatory environment, the professional-responsibility floor, the framework choice, the structural ownership, the policy stack, and vendor diligence. They are written to be useful on their own; readers already operating a mature program may skim them and move directly to Part II.

1. Why AI governance matters now

Adoption has outrun policy. The ACC and Everlaw 2025 GenAI Survey put active generative-AI use in corporate legal departments at 52 percent, more than double the prior year[1]. Clio reports majority adoption across the profession with roughly 44 percent of firms still operating without a formal AI policy[2]. The 2025 FTI Technology General Counsel Report found only 15 percent of GCs feel prepared to manage AI risk[3].

The regulatory calendar is no longer hypothetical. The EU AI Act prohibited practices applied in February 2025; general-purpose AI obligations applied in August 2025; most remaining provisions and enforcement powers apply in August 2026[4]. The Colorado AI Act, deferred to June 30, 2026 by SB25B-004[5], is being reworked by a policy work group that proposed an updated framework in March 2026[6].

Clients are pushing AI requirements into outside counsel guidelines: notice before material AI use, approved-tool lists, prohibitions on training vendor models with client data, audit rights, incident reporting, and in some cases billing limits. Sixty-four percent of in-house teams in the ACC and Everlaw survey expected to rely less on outside counsel as they built internal AI capability[1]. Firms that cannot answer governance questions are losing RFPs on that basis.

> **The one-sentence version**
>
> Adoption has outrun policy, regulation is arriving on schedule, and clients are asking governance questions that firms cannot credibly decline to answer.

Three forces converging

These three pressures (lawyer adoption, regulator timing, and client demand) are independent. A program designed only for one will buckle under the others. A program designed for all three is what this primer tries to help you build.

Anchor framework

Anchor governance to the NIST AI Risk Management Framework, organized around Govern, Map, Measure, and Manage, with the Generative AI Profile (NIST AI 600-1) as the operational supplement. Treat ISO/IEC 42001 as an optional certifiable overlay. Treat the EU AI Act, Colorado AI Act, and sector rules as binding obligations that map into NIST, not separate regimes you maintain on the side. The architecture stays simple, and obligations get added or retired without restructuring the program.

2. Professional responsibility baseline

Before any framework, the ABA Model Rules and their state analogs supply the floor. ABA Formal Opinion 512 (July 29, 2024) was the ABA's first formal opinion on generative AI[7]. It does not create new duties; it clarifies how existing ones apply.

The six rules that matter most

Rule	Duty	What Opinion 512 says, in brief
1.1	Competence	Lawyers need not be AI experts but must understand capabilities and limitations. Uncritical reliance on output without independent verification can violate the duty.
1.6	Confidentiality	Do not input client information into a generative AI tool without adequate protections. Boilerplate consent in engagement letters is inadequate. Get informed consent before client content enters a tool.
1.4	Communication	Tell the client when AI use materially affects how the representation is delivered, including cost or outcome.
3.3	Candor to the tribunal	The lawyer is responsible for every citation and representation. Fabricated authorities will be attributed to the lawyer, not the tool.
5.1 / 5.3	Supervision	Partners and managers must ensure firm policies and training produce reasonable compliance by lawyers and non-lawyers.
1.5	Reasonable fees	A lawyer may not bill for time not actually spent. Efficiency from AI may need to be reflected in the fee depending on the arrangement.

State and cross-border alignment

State bars have followed a consistent line: California Practical Guidance (Nov. 2023)[8], Florida Op. 24-1, NYC Bar Formal Op. 2024-5, North Carolina 2024 FEO 1, and Oregon Formal Op. 2025-205 each track Opinion 512 with local accents[9]. The Justia and NYC Bar 50-state surveys are the best current references[10]. For cross-border practices, the UK Solicitors Regulation Authority, the Bar Council of England and Wales, and several EU national bars have parallel guidance worth tracking[11].

Court rules and sanctions

Judge Brantley Starr's May 2023 standing order in the Northern District of Texas was the first to require disclosure and verification of AI-assisted filings[12]. Dozens of federal and state judges have since followed,

including Judge Michael Baylson of the Eastern District of Pennsylvania[13]. There is no FRCP-level rule yet; variation across judges is itself a governance challenge for national practices.

The sanctions record is now substantial. Mata v. Avianca, S.D.N.Y. 2023, imposed $5,000 in sanctions[14]; Park v. Kim, 91 F.4th 610 (2d Cir. 2024), referred an attorney for discipline[15]. Public trackers maintained by Damien Charlotin, Law360 Pulse, and the ABA Litigation Section catalog hundreds of incidents[16]. Treat the cases as motivation, not as the core of the controls themselves.

Operating checklist (translates duties into daily practice)

Independent verification of every citation, quotation, and factual claim before AI-assisted work product leaves the firm.

No input of client information into non-approved tools. Approved tools mean enterprise-tier deployments with contractual confidentiality and no training on client content.

Informed client consent (not boilerplate) before client content enters a tool, with the consent documented.

Designated supervising attorneys for each practice or department; training in onboarding and annual recertification.

AI-driven efficiency tracked and applied to the fee where the arrangement requires it.

3. External frameworks at a legal-industry level of detail

NIST AI Risk Management Framework and Generative AI Profile

The NIST AI RMF 1.0 (January 2023) organizes the work into four functions[17]. Govern is the leadership and policy layer. Map establishes context and inventories systems. Measure is the assessment layer, including testing and evaluation. Manage covers prioritization, treatment, and incident response. The Generative AI Profile (NIST AI 600-1, July 2024) names the risks lawyers recognize immediately (confabulation, data privacy exposure, intellectual property leakage, harmful bias) and maps mitigations back to the four functions[18].

ISO/IEC 42001:2023

ISO/IEC 42001 is the first international management system standard for AI, structured on Plan-Do-Check-Act familiar from ISO 27001[19]. Unlike NIST AI RMF, it is certifiable: an independent registrar can audit an organization against it. Pursue certification only when there is a credible business case (enterprise client demand, a regulated industry presence, or a third-party assurance need). For most legal organizations, conformance without certification is the right place to start.

EU AI Act

The EU AI Act is the first comprehensive, binding AI regulation, built around four risk tiers (unacceptable, high, limited, minimal) plus a distinct general-purpose AI regime. Its extraterritorial reach means a U.S. firm or in-house team with EU-facing systems, EU data subjects, or EU-placed services can be in scope. Key dates: prohibited practices February 2, 2025; general-purpose AI model obligations August 2, 2025; most remaining provisions and Commission enforcement powers August 2, 2026.

Colorado AI Act and the U.S. state landscape

Colorado SB24-205 was the first U.S. state law imposing affirmative duties on developers and deployers of high-risk AI systems. Its effective date was deferred to June 30, 2026 by SB25B-004, and the policy work group's March 2026 proposed framework may replace the original text. Other states are moving on a recurring pattern of risk tiering, impact assessments, transparency obligations, and recordkeeping. Anchoring to NIST and mapping each statute into that structure has worked better in practice than tracking each statute as a separate regime.

How to choose the anchor framework

For most law firms and in-house teams, anchor to NIST AI RMF and its Generative AI Profile.

Pursue ISO/IEC 42001 only when the business case is concrete and credible.

Treat the EU AI Act, Colorado AI Act, and sector rules as binding obligations mapped into the NIST architecture, not as alternative frameworks.

4. Building the governance program: structure and ownership

A governance program without named accountable leaders is a policy binder. The mature programs share a recognizable structure: an executive sponsor, a cross-functional committee, clear escalation, and working groups that do the work between committee meetings. A solo or small practice runs the same plays with fewer chairs in the room. The roles do not disappear; one person may hold several of them.

Scale it to your practice

Sizing the program to the practice

Solo or two to five lawyers. You are the executive sponsor, the committee, and the ethics lead. A one-page policy, a simple tool list, verification of every citation, and a twice-yearly review is a defensible program at your scale.

Small firm (6 to 40 lawyers). Name a partner as AI lead; add IT or your managed-services partner as advisor. Quarterly committee meetings are enough.

Mid-sized firm or in-house team (40 to 200). Stand up the full committee, but keep it lean. Rotate practice-group or business-unit representatives instead of adding permanent seats.

Large firm or enterprise legal department. The full structure maps directly: more sub-working groups, clearer RACI lines, and a reporting cadence to firm or board leadership.

Membership, charter, and decision rights

Standing seats: Legal or Office of the General Counsel (ethics and professional responsibility), Privacy, Information Security, IT or Engineering, Procurement or Contracts, and Human Resources. A rotating practice-group or business-unit sponsor keeps the committee close to actual use cases. The executive sponsor (managing partner or general counsel) does not need to attend every meeting but must own the charter and the escalation path. The committee approves Tier-high use cases by majority, with Legal holding a blocking vote on professional-responsibility questions.

RACI at a glance

Activity	Legal	Privacy	InfoSec	IT	Proc.	HR
AI policy and ethics review	R/A	C	C	I	I	C
Vendor due diligence	C	R	R	R	A	I
Client consent and engagement letters	R/A	C	I	I	I	I
Training and certification	C	C	C	I	I	R/A
Model and system inventory	C	C	R	R/A	I	I
Incident response	R/A	R	R	R	I	C

5. The policy stack and risk tiering

Most legal organizations need five written artifacts. Treat the list as a menu, not a stack of binders. A solo or small practice can combine the first two; a larger firm will want each as its own document. Full templates are best generated against the firm's actual operating model and reviewed by ethics counsel; this chapter focuses on what each artifact does and how the artifacts work together rather than reproducing one-size templates.

The five artifacts in one paragraph each

AI use policy. The top-level statement: permitted purposes, prohibited uses, who approves what, the escalation path. Short enough that lawyers actually read it.

Acceptable-use standard. Operational dos and donts by tier: approved tools, data types permitted in each, mandatory human review for client-facing output, prohibited use cases, reporting obligations.

Client-matter addendum (firms) or outside counsel guideline clauses (in-house). Engagement-letter language for client consent, categories of information not input into tools, the verification commitment, and billing treatment of AI-driven efficiency. Align to Opinion 512's rejection of boilerplate consent.

Vendor and model inventory. The living register of every AI tool in use, including informally-adopted free or individual purchases. The single document most frequently requested in client diligence and regulator inquiries.

Training and certification policy. Annual mandatory training tiered to role, with completion gated to system access where feasible.

The three-tier risk model

Tier	Scope	Controls
High	Client content, confidential matters, regulated data, consequential decisions about people.	Committee approval; DPIA or impact assessment; privilege analysis; mandatory human in the loop; audit log.
Elevated	Non-privileged drafting assistance, research, internal document analysis.	Approved tool list; standard vendor diligence; training required; human review before external use.
Low	Productivity and formatting, non-substantive tasks, public information.	Acceptable-use compliance; default tool settings; no PII or client content.

Tiering is the mechanism that keeps governance usable. It prevents low-impact uses from being smothered and high-impact uses from slipping through. A use case enters through an intake form that captures data types, population affected, decision autonomy, and deployment surface; the intake drives the tier assignment.

6. Vendor diligence and privilege

Most AI risk for legal organizations flows through vendors. A defensible diligence process is the highest-leverage control in the program. The goal is to know what the vendor does with firm or client content, document that knowledge, and backstop it contractually.

Security baseline

Require a SOC 2 Type 2 attestation for any tool handling client or confidential information. Where feasible, prefer vendors with ISO 27001 or ISO 42001 certification. Map sector-specific controls where the firm or client is regulated (HIPAA, GLBA, FedRAMP).

Contract terms that earn their keep

- Training data. Prohibit use of firm, client, or personal data to train, fine-tune, or improve vendor models. No exceptions for aggregated or de-identified data unless expressly negotiated.
- Confidentiality. Coverage of prompts, outputs, and logs at least as strong as the firm's standard NDA.
- Subprocessors. List with notice and objection rights; flow-down of all applicable obligations.
- Data location, retention, and deletion. Address client and regulatory locality requirements; defined retention windows; deletion on request and on termination.
- Audit rights. SOC report delivery, security questionnaire response, on-site or remote audit for high-tier uses.
- Indemnity. Intellectual-property indemnity covering training-data claims; privacy and security indemnity for breaches.
- Incident notice. Notice within 24 to 72 hours, with enough detail to meet the firm's reporting obligations.
- Model transparency. Disclosure of model family, material changes, and an evaluation summary for high-tier uses.

Privilege analysis

Disclosing client content to an AI vendor without adequate protections can waive attorney-client privilege. Courts ask four questions in some form: can vendor personnel read the content; does the vendor train on it; does an NDA strong enough to preserve privilege sit behind the arrangement; is the disclosure reasonably necessary to the representation. If you cannot answer yes on the privilege-preserving side to each, do not put client content through the tool.

Decision rule for the front line

Consumer tool, client content: never.

Enterprise tool on the approved list, client content: permitted with client consent recorded.

Any tool, fabricated citations: treat as a Rule 3.3 and 1.1 incident, not a benign mistake.

Part II

What governance actually requires

Part II addresses the harder questions that surface once the policy is written. Why do verification controls erode within a quarter? How should lawyers calibrate trust in a model whose confidence is, by construction, almost always high? When does it matter which model you used? Why does shadow AI persist even in firms that do everything right? What happens to associate development when the first draft of every memo comes from a machine? These are the questions that separate governance on paper from governance in practice.

7. Change management as the operating method

Rolling out AI governance is a change management problem before it is a documents problem. You are asking every lawyer in the firm or department to adjust how they draft, research, review, and bill. That kind of change takes eighteen months to land, not eighteen days, and it never lands at all if leadership treats the policy memo as the finish line.

This is the longest chapter in the primer because change management is where most programs fail in ways that the audit trail does not capture. The committee meets, the dashboard shows green, the training-completion number climbs, and lawyers quietly route around the system in ways no one will see until a sanctions order forces it into the open.

Why legal organizations are unusually hard to change

Three structural features make change management harder in law than in most knowledge-work settings, and each requires a deliberate counter-move.

First, the billable hour creates a perverse incentive against honest reporting of AI efficiency. A lawyer who finishes a brief in three hours instead of seven has a private interest in not advertising that fact. Firms that ignore this dynamic find that AI use is happening but is invisible to the governance program, which then cannot evaluate or improve it. The counter-move is to decouple the personal billing impact from the program's view: report AI use through a confidential channel that does not flow into individual billable-hour metrics, and separately decide how the firm wants to share efficiency with clients (some firms have moved to fixed-fee or value-based arrangements precisely to remove this friction).

Second, the partnership model spreads decision rights so widely that any policy is, in practice, a set of negotiations with several dozen people who each believe their practice is the exception. The counter-move is to identify and convert a small coalition of respected partners early, so that the message arrives from a peer rather than from the AI committee. Three to five visible champions, including at least one rainmaker, are worth more than a perfect policy.

Third, lawyers are professionally trained to spot risk and ask why a recommendation might be wrong. That makes them harder to win over, and also more careful when they do adopt. Treat the resistance as informed skepticism rather than obstruction. Answer the questions, including the ones that make the program look weaker, in writing. Lawyers who feel heard do not become evangelists; they become honest users.

Generational and role dynamics

Adoption splits in predictable ways. Associates and recent law school graduates are typically more comfortable with the tools and faster to find workarounds when policy gets in the way. Mid-level lawyers are the most enthusiastic adopters when AI helps with high-volume tasks like document review

or first drafts. Equity partners adopt slowest, often because they have the least time and the most reputational exposure.

Each group needs different framing. For associates, the message is verification rigor: AI does not lower the standard for what leaves the firm with their name on it, and the way to a partnership track is showing the same care with AI-assisted work as with anything else. For mid-level lawyers, the message is mastery: the lawyer who knows where the model fails on a contract review or a research task becomes more valuable, not less. For equity partners, the message is risk and revenue: clients are starting to ask, and the firms that can answer credibly will keep matters that the firms that cannot will lose.

This is also where the gendered dimension of change management deserves a sentence. In several published surveys, women in legal practice report higher concern about AI risk than their male colleagues and lower confidence that their firm has adequate controls. Whether or not those gaps reflect underlying differences in risk tolerance, they reflect a real difference in how the program is experienced. A change effort that does not actively check whether the message is landing across cohorts will quietly leave half the firm behind.

Partner buy-in: the practical playbook

Partner buy-in is the part of change management that programs most often try to skip. Three moves do most of the work.

Run a short pilot in a high-visibility practice. Pick a partner with credibility and a use case where the model clearly earns its keep, such as deposition summarization, large-document review, or initial NDA triage. Define one or two measurable outcomes (time saved, accuracy holding, client satisfaction unchanged) before the pilot starts. Publish results, including any near misses, when it ends. A six-week pilot in a single practice produces more partner buy-in than six months of committee meetings.

Bring partners into the failure stories before the failure stories happen to them. Walk through the Mata v. Avianca pattern, the Park v. Kim pattern, the recent state-court hallucination orders. Walk through a near miss from inside the firm, anonymized. Lawyers respond to specifics. Abstract risk discussions slide off.

Make the policy partner-respecting. The policy should make life easier for a careful partner, not harder. If the approved-tool list adds friction without obvious benefit, partners will route around it. Show the path: here is the tool, here is the workflow, here is what we already negotiated for confidentiality, here is the time you save by not having to triage on your own.

Communication that treats lawyers as adults

Legal organizations consistently underinvest in the communication side of change. The implicit theory is that smart professionals will read the policy and comply. They will not. They will read the policy, decide which parts make sense, and then update those parts in their head over time as the situation changes.

The best programs run a continuous, low-volume communication rhythm. A short monthly note from the executive sponsor that names a real problem the program solved, names a real near miss, and links to one updated reference. A quarterly anonymous survey, with the results published. A standing office hour that any lawyer can drop into. None of these are expensive; collectively they are the difference between a program that lawyers reference and one they tolerate.

The reporting climate

Most AI incidents surface because a lawyer notices a fabricated citation, realizes they pasted the wrong content into a tool, or sees a peer do something problematic. Those reports come in only when the culture rewards honest escalation. Build a no-blame channel for near misses, publish anonymized lessons learned, and make clear that the reportable offense is hiding an incident, not having one.

Be explicit about the reverse case as well. A lawyer who follows the policy, hits a hard edge, and reports it should be visibly thanked, not subjected to a process that punishes the visibility. The first time a careful junior lawyer gets pushback for raising a concern, the reporting climate is over for two years.

A practical change management rhythm

First six months

Months 1 to 2. Publish the case for change in a one-page note from the managing partner or general counsel. Name three to five champions. Open a confidential feedback inbox. Begin the inventory of informal tool use.

Months 3 to 4. Launch baseline training. Run one open session per practice group. Ship a plain-English FAQ that includes the awkward questions. Approve and announce two to three high-visibility pilots.

Months 5 to 6. Publish first metrics, including at least one anonymized near-miss story. Recognize a careful-use story by name. Refine policy based on feedback. Add the AI use question to the matter intake form.

After the first six months

Quarterly. One short written update from the governance committee. One live forum. One refinement based on what you learned.

Annually. A full review of the policy stack. A re-run of training, with material updated to reflect the

year's lessons. A presentation to the executive committee or board.

On every model upgrade. Bring the change to the committee. Communicate to lawyers what changed. Update training touchpoints if behavior shifts in ways that matter.

Cultural warning signs

Five signals that the program is in trouble even when the dashboard is green

Lawyers say I do not use AI when they mean I do not report it.

Training completion is high but no one can describe the approved tool list from memory.

The governance committee has not seen a near-miss report in six months.

Associates feel they cannot raise an ethics concern with a supervising partner.

Senior leaders delegate AI entirely to IT or Legal and do not engage themselves.

8. The verification burden

Verification is the single control that every state bar opinion, every standing order, and every honest practitioner agrees on. It is also the control most likely to fail in production. Not because lawyers reject the duty (they do not) but because the duty asks for a kind of work that, done thoroughly, often takes as long as the original task. That tension is the verification burden, and getting it wrong is the most common way a program that looks compliant on paper produces hallucination incidents in fact.

The productivity paradox in one paragraph

If a model produces a five-page memo in two minutes and a careful citation check, factual review, and reasoning audit takes three hours, the lawyer has saved nothing. If the same lawyer skips the verification, the firm has saved time on average and accepted catastrophic tail risk. Most lawyers, faced with this trade quietly each week, will compromise: a quick look, a spot check, a confidence in the model that grows with the number of times it has not embarrassed them. That compromise is how Mata v. Avianca happens to a lawyer who would have caught the same problem in a junior associate's draft[14].

What verification actually requires

Verification is not a single act. It is at least four distinct activities, each with different costs and different failure modes.

Citation verification. Confirming that every case, statute, regulation, and secondary source the model cited exists, says what the model says it says, and supports the proposition for which it is cited. The most familiar verification task and the one most often skipped under time pressure.

Quotation verification. Confirming that quoted material is accurate to the source. This includes block quotes, parentheticals, and inline phrasing presented as if drawn from a source. Models routinely paraphrase while presenting the result inside quotation marks.

Factual verification. Confirming that non-citation factual claims (dates, parties, holdings, procedural posture) are accurate. This is where models silently introduce errors that a citation check will not catch, because the citation may be real and the proposition wrong.

Reasoning audit. Confirming that the legal analysis follows from the cited authority, applies the right standard, and addresses the controlling jurisdiction. The most expensive form of verification, the one that requires the most senior judgment, and the one almost never delegated to a tool.

Why the verification burden grows over time

Three dynamics make verification harder in production than in early pilots.

First, model fluency rises faster than model accuracy. Outputs that look more authoritative are not more reliable; they are harder to second-guess. The same lawyer who carefully verified a clunky 2023 draft trusts a polished 2026 draft because it reads like work product. Polished prose is, in this respect, an active enemy of careful review.

Second, lawyers calibrate to the model's average behavior, not its tail behavior. After a model has produced ten correct citations in a row, the eleventh draft does not get the same scrutiny. The eleventh is when the fabrication appears. NIST calls this confabulation; the operational problem is that human verification habits are tuned to base rates, not to the kinds of failures that cost careers.

Third, the verification step is asymmetric in incentives: a lawyer who skips verification and gets away with it pays nothing; a lawyer who skips and gets caught pays everything. The asymmetry favors caution, but only weakly, and only at moments of clear visibility. In the steady state, where most work is invisible to anyone but the lawyer, the asymmetry quietly fades.

Designing verification so it survives in practice

Programs that hold up tend to share five design choices.

Make verification an integrated part of the workflow, not an additional task. Tools that surface citations linked to the underlying authority while the lawyer reads, that display each quoted phrase next to the actual source text, and that flag claims with no provenance reduce the cost of verification by an order of magnitude. The single most important question to ask a vendor of any legal AI tool is: what does verification look like inside this product?

Use retrieval-augmented generation (RAG) where citations matter. RAG-grounded systems retrieve real documents and generate responses from them. They still hallucinate, but at lower rates, and they make

verification cheaper because the retrieved sources are visible to the lawyer. Treat RAG as a meaningful risk reducer, not a guarantee.

Tier verification by stakes, not by tool. A research memo for an internal training is not a brief. The verification rule should match the consequence of being wrong. Make the tiering visible: tag every output with the tier the lawyer assigned, and review whether the assignment matched the use.

Pair every AI-assisted draft with a verification checklist that the lawyer initials before circulation. The checklist need not be long. It should name the four verification activities, ask the lawyer to confirm each, and require a one-line statement of the verification approach. The point is not the form; the point is the cognitive interrupt that makes the lawyer think the question through. A two-minute checklist routinely prevents a ten-thousand-dollar sanction.

Build automated checking into the pipeline. Citation lookup against Westlaw, Lexis, or Fastcase to confirm that cases exist and have not been overruled. Quotation checking against the cited source. Hallucination detection on cited propositions. Several commercial tools (Appendix A) do this; for many practices, the right answer is to layer one of these on top of the drafting tool.

When verification is enough

Lawyers ask, reasonably, when verification can stop. The honest answer is that the duty under Rule 1.1 is reasonable verification, not exhaustive verification, and reasonableness varies by stakes, audience, and the nature of the underlying work. A research memo for an internal partner update tolerates a different level of scrutiny than a Section 1983 complaint.

A working test: would a careful lawyer in your jurisdiction, knowing what you know about the tool and the matter, sign their name to this without further review? If the answer is no, the verification is not enough. If yes, document what verification was done, and move on. The documentation is what protects the lawyer if a question arises later. Saying I verified is not enough; saying I verified by checking each citation in Westlaw, comparing each quotation to the source, and reviewing the reasoning against [authority] is.

Five questions to ask before AI-assisted work product leaves the firm

Have I confirmed every cited authority exists, is good law, and supports the cited proposition?

Have I checked each quotation against the actual source, not against the model's representation of it?

Have I tested the factual claims that are not citation-bound (dates, parties, procedural posture, holdings)?

Have I reviewed the legal reasoning against the controlling authority for the relevant jurisdiction?

Would I be comfortable defending this verification, in writing, to a court or to bar counsel?

The honest version

If the verification step routinely makes AI assistance unprofitable for a given task, that is a signal that the task is not yet a good fit for AI assistance, or that the tool is not yet good enough at it. The mistake is to keep using the tool while quietly degrading the verification, because that is the path that ends in sanctions. Either upgrade the tool, find a better-fit task, or accept the time cost. Pretending the trade is not there is what programs that look fine on paper do, right up until they do not.

9. Trusting AI models: a working framework

Trust in an AI model is not a binary. It is a continuously updated estimate of how reliable the model is for a given task, with a given dataset, in a given moment. Lawyers, who are trained to evaluate witnesses, sources, and arguments by exactly this kind of running judgment, ought to be unusually well-equipped for it. In practice they often are not, because the surfaces a model presents (fluent prose, confident tone, plausible structure) do not vary with the underlying reliability.

This chapter sets out a working framework for calibrated trust in legal AI. It borrows from the literature on trust calibration and uncertainty quantification, but it is written for working lawyers rather than for researchers.

The two failure modes

Trust failures come in two shapes. Overtrust is the failure most lawyers worry about: relying on a model output that is wrong. The Mata v. Avianca pattern. Undertrust is the failure most lawyers do not notice: refusing to use a model output that is, for the relevant purpose, more reliable than the alternative. Spending eight hours on a research task the model would have done correctly in twenty minutes is a real cost, even if it never appears on a sanctions docket.

Both failures matter. A program that addresses only overtrust produces lawyers who use AI badly because they do not use it at all. A program that addresses only undertrust produces lawyers who get sanctioned. Calibrated trust is the middle; getting there requires giving lawyers a vocabulary for it.

A working vocabulary

Three concepts borrowed from the trust calibration literature, in plain English.

Reliability. The probability that the model will produce a correct output for the kind of task you are giving it, with the kind of inputs you are providing. Reliability is task-specific. A model that is highly reliable at summarizing depositions may be much less reliable at drafting jurisdictional analysis. Treat reliability as a per-task estimate, not a property of the model.

Calibration. The match between the model's apparent confidence and its actual reliability. A well-calibrated model is uncertain when it should be uncertain. Most current frontier models are poorly

calibrated by default: they sound confident on questions where they should not be. The lawyer's job is to do the calibration the model does not.

Epistemic humility. The recognition that you, the lawyer, also have limits in knowing when the model is wrong. A model can produce a plausible answer to a question whose correct answer requires expertise the lawyer lacks. The temptation to defer to the model when the lawyer cannot independently evaluate the output is the most subtle and most dangerous form of overtrust. The discipline is to refuse to use AI output for anything you cannot verify, even when the model sounds right.

Where models are reliable, and where they are not

As of mid-2026, frontier general-purpose models are reasonably reliable at: structured summarization of clearly-bounded source documents; first-pass drafting of routine documents (NDAs, demand letters, internal memos) using your own templates as inputs; identifying issues in a contract against a checklist; reformatting and translating; explaining concepts to non-experts.

They are unreliable, in ways that often look reliable, at: case research without retrieval grounding (free citation generation is the canonical hallucination case); jurisdictional analysis where the controlling authority is recent or rapidly evolving; multi-step legal reasoning that requires holding many constraints in mind at once; calibrated estimates of risk or likelihood; questions that require reading an unfamiliar tone or strategic intent.

Specialty legal models trained or grounded on legal corpora (Westlaw AI, Lexis+ AI, CoCounsel, Harvey, and others) typically reduce the hallucination rate on legal tasks by an order of magnitude relative to general-purpose models, but Stanford's 2024 study (Magesh et al.) found legal-specialty tools still hallucinate on 17 to 34 percent of queries depending on tool and task[20]. The lower base rate is real and useful; it does not eliminate the verification duty.

The calibration drill

Trust in a tool builds and erodes through experience. The lawyer who has used a model for three months has a richer estimate than the lawyer who read the brochure. Programs can accelerate that calibration deliberately.

Run a calibration exercise once a quarter. Pick five tasks the firm uses the model for routinely. For each, have a senior lawyer compare model output against a known-good answer. Score each on three axes: accuracy, completeness, and confidence-calibration (did the model express uncertainty where it should). Publish the results internally with examples. After three or four quarters, the firm has an actual reliability map for the tools it uses, not a vendor brochure.

Make the failures discussable. The fastest way to build calibrated trust is to expose lawyers to specific, concrete cases of model failure on tasks like the ones they do. A monthly five-minute slot in the practice meeting for a recent failure example does more for calibration than an hour of policy discussion.

When the model and the lawyer disagree

A subtle but common situation: the lawyer drafts something one way; the model suggests a different approach; the lawyer is not sure which is right. The default should be the lawyer, especially in adversarial contexts where the model has no view of strategy or relationship. The model is a generalist working from training data; the lawyer has the matter, the client, and the jurisdiction in front of them.

If the model's suggestion seems clearly better, treat that as a signal to verify the underlying authority before adopting it. If the model and the lawyer reach different answers on a research question, the lawyer's answer should be the working hypothesis until the model's answer is verified. This is the opposite of how many lawyers, especially junior ones, instinctively work. The instinct should be deliberately reversed.

> **A trust calibration heuristic**
>
> If you cannot independently evaluate the output, do not use it.
>
> If you can evaluate it but have not yet, treat it as a hypothesis, not an answer.
>
> If you have evaluated it and it holds, document what you did. The documentation is the bridge between trust and accountability.
>
> Reset the calibration whenever the model is upgraded. Vendor major-version changes can shift behavior in ways that invalidate prior trust estimates.

10. Model differences and what changes when you switch

Most AI governance discussions treat the model as a black box. That works at the policy level. It breaks down operationally, because frontier models behave differently in ways that matter for legal work. A program that does not understand those differences will make procurement choices that look identical on paper and produce different outcomes in practice, and will respond to model upgrades as if they were minor when they are sometimes substantial.

This chapter is a working overview of how the major model families differ as of mid-2026, what those differences mean for legal workflows, and how to build the difference into the governance program.

The frontier models in one paragraph each

Anthropic Claude (Opus, Sonnet, Haiku families). Strong on long-document analysis, careful structured drafting, and nuanced legal reasoning. Tends to express more uncertainty than peer models, which is helpful for verification but can read as hedging in client-facing output. Anthropic's commitments around training data and refusal behavior are stricter than peers, which lawyers tend to value. Frequently rated highest by working lawyers for legal drafting quality and IRAC-style reasoning.

OpenAI GPT (4 / 4o / 5 families). Broadest tool ecosystem and the largest base of legal-specific add-ons. Strong general performance, with a richer plug-in surface (Custom GPTs, function-calling, broader integrations) that suits firms building internal workflows on top of the model. Web search and browsing add real-time research capability. Has been most associated with high-profile hallucination incidents, partly because it is the most-used model rather than because it is the most error-prone.

Google Gemini (1.5 / 2.x families). The largest practical context window, which matters when feeding entire case files, transcripts, or document productions into a single prompt. Native Google Search grounding gives current-events advantage for research that turns on recent regulatory developments. Multimodal capability (handling images, audio, and documents in the same prompt) is well ahead of peers, which matters for practices doing visual evidence review.

Specialty legal models (Harvey, CoCounsel, Westlaw AI, Lexis+ AI, Paxton, Spellbook, Legora, and a growing list). Built on top of one or more frontier models, with retrieval grounding, evaluation pipelines, and jurisdiction-aware features layered on. Hallucination rates are meaningfully lower on legal tasks than the underlying frontier models alone. The Stanford Magesh study (2024) gave legal-specialty tools 66 to 83 percent accuracy on legal questions, against 12 to 31 percent for general-purpose models on the same questions[20].

How models actually differ

Six dimensions account for most of the practical differences.

Reasoning depth. How the model handles multi-step legal analysis where each step depends on the previous one. The honest answer is that all frontier models still degrade as the analysis chain gets longer, with degradation rates that vary by family. For legal work, this is the single most important dimension and the one the marketing materials disclose least.

Citation behavior. Whether the model invents citations, retrieves them, or refuses when it cannot ground the citation. Frontier models without retrieval will fabricate; models with retrieval will sometimes still fabricate but at much lower rates. The right question is not does the model hallucinate but what does the model do when it does not know.

Refusal posture. How the model behaves when asked to do something it should not do, or when a prompt is ambiguous. Stricter refusal saves the lawyer from misuse but creates friction; looser refusal is more cooperative but lets misuse through. Anthropic, OpenAI, and Google have made different trade-offs here that lawyers should understand for the workflows they automate.

Confidentiality posture. Whether and how the vendor uses prompts to train or improve models. Enterprise tiers from major vendors generally commit to no training on customer content, but the contractual fine print varies. This is the contractual question that most directly engages Rule 1.6, and it should be answered the same way for every approved tool: in writing, with a named signatory.

Context window. How much information the model can hold in a single conversation or prompt. Larger windows allow more documents in a single prompt; they do not, by themselves, mean better reasoning across the documents. A 1-million-token window is useful when the input genuinely exceeds smaller windows; otherwise it is mostly marketing.

Update cadence. How often the underlying model changes and how the vendor handles change communication. A model that changes silently is a governance problem; a model that ships changes with notes, evaluations, and migration guidance is governable. This is the single most underweighted factor in vendor selection.

A practical model-selection matrix

Task	Model strengths	Verification implication
Long-document review (deps, productions)	Gemini for raw context size; Claude for nuanced reading	Spot-check sampling across the document; do not assume completeness from a single pass.
Contract drafting and redlining	Claude or specialty tools (Harvey, Spellbook, CoCounsel) for structured drafting	Compare to your standard playbook; treat the model's positions as a starting draft, not a redline.
Legal research	Specialty tools (Westlaw AI, Lexis+ AI, CoCounsel) with grounded retrieval	Verify every citation in the underlying database, even with retrieval grounding.
Current events / regulatory updates	Gemini or GPT with web grounding	Confirm against the source publication; AI summary may misstate the timing or scope.
Internal summarization / first drafts	Any frontier model on enterprise tier	Treat as draft only; lawyer rewrites for client-facing posture.
Strategic analysis (settlement, risk, story)	Lawyer-led, model-assisted at most	AI lacks matter context; use as a counter-perspective, not as the answer.

Multi-model approaches and orchestration

A growing number of legal AI products do not commit to a single underlying model. They route requests across two or more frontier models depending on the task, sometimes with a proprietary model for

sensitive workloads. Thomson Reuters CoCounsel, Harvey, and several others have publicly described this orchestration pattern.

For governance, multi-model architectures change the question from which model are we using to which model is being used right now and on what basis. The committee should know, for any approved tool, which models can be invoked, how the routing decision is made, and what changes when the vendor adds a new model. Insist on that disclosure in writing during procurement; many vendors will treat it as table stakes once asked.

What to do when a model upgrades

Vendor model upgrades have, in several documented cases, materially shifted output behavior. Refusal rates change. Citation behavior changes. Tone changes. Workflows that depended on prior behavior break in ways that are hard to predict from the release notes.

A working response: every vendor with an approved tool gets a designated owner on the governance committee. When a major version arrives, the owner reviews the release notes, runs a small re-evaluation against the firm's known-good tasks, and reports to the committee whether anything has changed enough to update training, the policy, or the tier assignment. For most upgrades the answer will be no; the cases where it is yes are exactly the ones the program needs to catch early.

Open weights and self-hosted considerations

A subset of legal organizations, particularly larger in-house teams or specialized practices, are evaluating open-weights or self-hosted deployments (Llama, Mistral, and others) for confidentiality reasons. The trade is real: data never leaves the firm's infrastructure, but the firm assumes responsibility for hosting, monitoring, evaluation, and update cadence that a hosted vendor would otherwise carry.

For most firms, the math does not yet work; the operational overhead is greater than the marginal confidentiality benefit over a well-negotiated enterprise contract with a hosted provider. For some, particularly those handling national-security or highly regulated work, the math does work. The committee should treat self-hosted deployment as its own tier, with its own controls, rather than assimilating it to the standard vendor model.

11. Shadow AI, skill atrophy, and other underrated risks

Three risks tend to receive less attention than hallucination, vendor breach, and confidentiality, even though each can quietly undermine a program over time. This chapter takes them in turn.

Shadow AI

Shadow AI is the use of unsanctioned AI tools by lawyers and staff inside the firm. Recent surveys put the gap between actual AI use and firm-provided tooling at twenty to thirty percentage points, meaning a substantial share of lawyers are using tools the firm has not vetted, on data the firm has not authorized, with controls the firm cannot enforce[21].

Shadow AI is rarely an act of bad faith. It is most commonly the response to a tool being unavailable, slow, or worse than the consumer alternative the lawyer can use on their personal device. The honest read is that shadow AI is a market signal: lawyers are telling the firm what they need, in the form of behavior the firm did not authorize. Treat it that way.

Three controls usually work better than the alternatives. First, make the approved tool genuinely better than the consumer alternative for the work that matters; investment here pays back across the program. Second, run an amnesty inventory once a year that asks every lawyer to disclose, without consequence, any tool they have used that is not on the list; populate the inventory from honest answers and update the approved list to reflect what people actually need. Third, deploy network or endpoint controls (DLP, browser extensions, or governance platforms like TruthSystems and the platforms named in Appendix A) that block disclosure of client content to unsanctioned services without requiring the lawyer to remember what is on the list. These controls work best in combination; none alone is sufficient.

Skill atrophy in junior lawyers

If the first draft of every research memo, contract, and discovery analysis comes from a model, the junior lawyer who would historically have written that first draft does not develop the skill that comes from doing so. Within a few years, this becomes a structural problem: the firm has senior lawyers who can review AI output and junior lawyers who cannot, and no clear path from one to the other.

There is no consensus solution yet. The approaches that look most promising do three things. They preserve a meaningful share of foundational drafting work without AI assistance, especially in the first two years of practice. They make AI use visible in the development conversation, so that mentors can coach junior lawyers on judgment and verification rather than only on output. And they redefine what good associate development looks like to include AI literacy as a tracked skill, not as a side project.

This is also a recruiting and retention question. Junior lawyers who feel they are being deprived of foundational training will leave. Junior lawyers who feel they are being asked to do AI-augmented work without the support to do it well will also leave. The answer is not to ban AI for first-year associates and not to let it replace foundational training; it is to integrate the two deliberately.

Vendor lock-in and exit planning

Legal AI tools often integrate deeply with firm systems: matter management, document management, billing, client portals. The deeper the integration, the harder it becomes to switch vendors when the contract terms change, a competitor offers a better product, or the vendor degrades the service after acquisition or pivot.

Build exit considerations into procurement from day one. Contractual rights to data export in usable formats. Documented portability of any custom configurations, prompts, or workflows. A migration plan that the firm can execute without vendor cooperation if it has to. None of this is glamorous; all of it pays off the first time a tool the firm relied on quietly stops being the right answer.

Cross-jurisdictional behavior differences

Frontier models perform differently on questions of U.S. law than on questions of UK, EU, or commonwealth law, and even within the U.S. their reliability varies sharply by jurisdiction. Models trained primarily on American legal corpora can be confidently wrong on Quebec civil law, Scottish procedure, or specific state regulations, in ways that are not obvious from the output.

For cross-border practices, the working rule is to assume the model is less reliable on jurisdictions outside its training-data center of mass, and to require local-counsel verification on matters that turn on those jurisdictions. The same caution applies to specialized U.S. domains (tribal law, military justice, certain regulatory regimes) where training data is thinner.

Bias and disparate impact in legal use

AI bias gets the most attention in hiring, lending, and criminal-justice contexts, but it shows up in legal practice in subtler ways. Models can make systematically different recommendations on settlement ranges, sentencing predictions, or risk assessments based on demographic features in the inputs. They can produce different drafting tone for parties with names that read as different ethnicities. They can summarize witness statements with different emphasis depending on the witness's apparent identity.

For most law firms, the right response is not algorithmic auditing of every tool (that is the vendor's job) but operational sensitivity: include bias-relevant checks in the verification step for any AI use that materially affects how the firm treats people, particularly in employment, housing, criminal defense, and family law. Document the checks. Be ready to discuss them with clients who ask, because in 2026 some clients will ask.

12. Insurance, billing economics, and liability allocation

AI use changes the economics of legal work and the allocation of risk in ways that the policy stack alone does not address. This chapter covers three operational areas where the change shows up directly:

professional-liability insurance, billing model implications, and how liability gets allocated when something goes wrong.

Professional liability and AI

Major legal-malpractice insurers have started asking AI questions on renewal applications in 2025 and 2026: do you have a written AI policy; do you have a tool inventory; do you provide training; do you require verification of AI-assisted work product. Answers are starting to affect both pricing and coverage availability. Gartner has publicly recommended that general counsel evaluate AI-specific insurance products as part of risk management[22].

The honest summary as of mid-2026: standard professional-liability policies likely cover sanctions and malpractice claims arising from AI use, but the coverage analysis is not always clean. Some policies have introduced AI-specific exclusions or sub-limits. Some have added AI-specific endorsements with their own conditions. The policy at renewal is meaningfully different from the policy of three years ago, and it is worth a careful read with broker and coverage counsel.

Two practical moves. Get the firm's current policy reviewed for AI-related language by someone who reads coverage policies for a living; assume the form has changed. And keep the documentation that the carrier will ask for after an incident: the policy in effect at the time of use, the training records, the tool approval history, the verification record on the specific matter. Carriers reward the firms that can produce this. Carriers price punitively against firms that cannot.

Billing economics: who keeps the savings

AI-assisted work can reduce the time required for a task by an order of magnitude. That creates a contractual and ethical question about who captures the savings, and the honest answer is that it depends on the fee arrangement and the client's expectations.

Hourly billing. Rule 1.5 forbids billing for time not spent. AI-assisted work that takes less time can be billed for less time, full stop. Where the hourly rate already reflects the lawyer's expertise, that may be the end of the analysis. But many clients now expect that the firm will share efficiency gains beyond the simple time reduction, particularly when the AI tool was paid for by the firm and the savings flow to the firm. Be ready to discuss this; do not be the firm that learns about the expectation by reading it in an outside counsel guideline.

Fixed fee. AI-driven efficiency typically increases the firm's margin on fixed-fee work without contractual implications, which is part of why fixed-fee arrangements are growing in popularity. The risk is that clients begin to demand fixed fees that assume AI efficiency, eroding the margin to zero. Price the work, not the hours.

Value-based and contingency. AI does not change the analytical structure of these arrangements but it can change the firm's cost structure underneath them. Track the cost shift; renegotiate where the original arrangement no longer reflects the underlying economics.

Liability allocation: firm, vendor, lawyer

When an AI-assisted error reaches a client or a court, three potential targets for liability exist: the firm, the vendor, and the individual lawyer. How responsibility is allocated among them depends on the contract terms, the firm's controls, and the specifics of the incident.

The vendor contract typically caps the vendor's liability at the fees paid, with carve-outs for confidentiality breaches and intellectual-property infringement that may run higher. The firm's professional-liability carrier covers the firm and the lawyer, subject to the policy's terms. The lawyer's bar discipline is independent of any of this and is the most common adverse outcome in current AI-incident cases.

Three implications. First, do not assume the vendor will be the primary source of recovery; the contractual cap usually makes that economically marginal. Second, the firm's investment in controls and verification is what mostly determines whether the firm and the lawyer are exposed; pay for the controls. Third, document everything, because in a discipline proceeding the available record is what determines the outcome more often than the underlying facts.

Documentation that pays back the most under stress

The approved tool list and tier assignment in effect on the date of the incident.

Training completion records for the lawyer and any supervising attorneys.

The verification record on the specific work product, including the checklist if one was used.

The vendor contract in effect at the time, including the data-handling and incident-notice terms.

Any communication with the client about AI use on the matter.

13. Agentic systems and the next governance frontier

Most of this primer addresses AI systems that respond to a single prompt with a single output, with a human in the loop for every meaningful action. That model is changing. Agentic systems (AI that takes multi-step actions across tools, sometimes including external systems, sometimes including the firm's own infrastructure) are moving rapidly from research demos to production deployments in late 2025 and through 2026.

This chapter is shorter than the others because the law and the practice are both still settling. The point is to flag the dimensions a governance program should be ready for, not to prescribe controls that may be obsolete in eighteen months.

What changes when an agent acts on its own

Three things change when AI moves from response to action.

Verification becomes harder. A single output can be reviewed before it leaves the firm. A multi-step action sequence (open the contract, edit the term, send the redline to opposing counsel, file the calendar entry) is harder to interrupt mid-stream and easy to skip reviewing in the moment. The verification regime designed for response-only AI does not transfer cleanly.

Authorization becomes a real question. When a model writes a draft, the lawyer who edits and sends it is the actor of record. When an agent takes an action against an external system, the question of who authorized the action becomes contested. Vendor terms increasingly try to push that responsibility to the user; firm controls have to assume that and design accordingly.

The blast radius grows. A response-only model produces text. An agent can produce text, send it, schedule follow-on work, modify firm records, and trigger third-party systems. The cost of an error scales with the action's reach; a hallucinated email sent automatically to opposing counsel is not the same as a hallucinated email a paralegal could have caught.

Initial governance moves

- Treat agentic deployments as their own risk tier above high. Default to off; require explicit committee approval for any deployment that takes external action; require named owners and pre-approved scopes.
- Limit scope tightly. An agent that can read documents and propose actions is governable; an agent that can read, decide, and act unsupervised is, in most legal contexts, not yet ready for production.
- Require audit logs that capture every action the agent took, the inputs that triggered the action, and the human approval (or pre-approval) that authorized it. Treat the logs as records subject to the firm's retention policy.
- Build a circuit breaker. Every agentic system should have a documented procedure for halting it on demand, and that procedure should be tested at least quarterly.
- Update outside counsel guidelines. Most current OCGs were not drafted with autonomous action in mind; clients will start asking soon. Be ready with a position before the question arrives.

Agentic AI is the area where the governance program built for 2024 will most clearly show its limits. The investment is in keeping the program loose enough to absorb the change without rebuilding from scratch.

Appendix A. Researched tools landscape

This appendix profiles a small set of commercial offerings that show how the ideas in this primer translate into shipping products. Inclusion is illustrative, not an endorsement. Pricing, features, and positioning move quickly; verify directly with the vendor before any procurement decision. Each entry was researched against vendor materials and third-party coverage at the time of writing.

Governance and runtime control platforms

TruthSystems (truthsystems.ai)

TruthSystems positions itself as infrastructure for legal AI risk management. Its product, branded Charter, transforms a firm's written AI policies into runtime guardrails: blocking prompts that would violate confidentiality or policy, dynamically scoping which tools an attorney can access based on the matter at hand, and capturing granular audit trails of AI tool interactions across the firm[23]. The pitch is to make policy enforceable at the moment of use rather than after the fact, which is the place most policies fail.

For legal teams, TruthSystems sits in the gap between an acceptable-use policy and a DLP product: it understands the legal context (matters, clients, conflicts) better than a general-purpose data-loss tool, and acts on rules a firm has actually written. Reported features as of 2026 include browser-extension and platform deployment options, real-time prompt screening, and audit-log export for compliance. The company secured a reported $58 million funding round in January 2026, which signals market traction in this category but does not by itself attest to product fit[24].

Practical evaluation questions for a legal buyer: how does the product reconcile firm policy with vendor terms when they conflict; what is the false-positive rate for prompt blocking on legitimate work; what integrations exist with the firm's identity and matter-management systems; can the audit logs be exported in formats the firm's record-retention system can ingest.

WitnessAI (witness.ai)

WitnessAI markets a unified AI security and governance platform spanning employees, models, applications, and emerging agents. Reported capabilities include real-time data redaction on prompts and outputs, behavioral controls keyed to user intent, monitoring for drift and bias, and compliance support including the EU AI Act. WitnessAI has been used as a comparison point in trade-press coverage of the runtime-controls category[25].

Evaluation considerations are similar to TruthSystems: the tool is useful insofar as it enforces rules the firm has actually defined, integrates with the firm's identity stack, and produces logs the firm can use after an incident.

FairNow (fairnow.ai)

FairNow markets an AI governance platform that centralizes inventory, automates bias testing and transparency documentation, and maps use cases to regulatory frameworks including the EU AI Act, ISO/IEC 42001, and NIST AI RMF. Reported features include role-based workflow assignments, automated alerts on deadlines and new risks, and integration with MLOps, GRC, and ticketing tools[26]. The UK Government AI Assurance Techniques catalog lists FairNow as an implementation option[27].

FairNow is more oriented to inventory and assessment than to runtime control; legal teams typically evaluate platforms in this category for the quality of regulation-to-control mapping and the usability of the assessment templates.

Holistic AI (holisticai.com)

Holistic AI is a broader-market enterprise AI governance platform offering full-lifecycle oversight from model discovery through risk management and compliance[28]. It is not legal-specific; it is the platform several large enterprise legal teams evaluate when they need governance that scales beyond the legal department to the rest of the business. For in-house teams in regulated industries, this category is increasingly part of the evaluation set.

CounselGuard (counselguard.io)

CounselGuard is a purpose-built AI governance and compliance platform for law firms, positioned as the compliance system of record for AI in legal practice[29]. The pitch is to replace the spreadsheets and shared documents firms currently use to track AI use with a continuous, auditable platform that maps tool use to specific ethical obligations across jurisdictions, matters, and users. As of this revision the product is in early-access onboarding for law firms via waitlist.

The platform is organized around a small set of modules visible in the main navigation: Dashboard, AI Tools, Compliance, Policies, Training, Activity, Reports, and Audit Log. The functional emphasis tracks the operational gaps this primer flags throughout: knowing what tools are in use, mapping that use to the right rules, keeping the policy stack current, evidencing training, and producing the audit trail an inquiry will demand.

Key capabilities the vendor describes:

- Jurisdiction-aware compliance. Tracks obligations across ABA Formal Opinion 512 (US), the EU AI Act, SRA Standards (UK), Canadian law society guidance (FLSC Model Code), and Australian conduct rules, with per-jurisdiction status visible at a glance.
- AI tool registry. Maintains a catalog of every AI tool in use across the firm, with vendor assessments. The illustrated registry includes Harvey, CoCounsel, Microsoft Copilot, ChatGPT, Claude, Gemini, Perplexity, Lexis+ AI, Spellbook, and Kira, which is a fair reflection of what most firms actually have in production.

- Audit-ready reports. Generates compliance evidence packages for bar audits or regulatory inquiries on demand, addressing the documentation problem this primer treats as a leading indicator of program health.
- Policy management. Provides version-controlled policies with timestamps, approvals, and staleness tracking for AI usage policies, data handling, and vendor assessment standards. The staleness tracking matters: most firms do not have a reliable signal that the policy has aged out of its supporting authority.
- Training tracking. Monitors completion rates by practice group and flags overdue training obligations.
- AI usage monitoring. Creates immutable audit trails of who used which tool, when, and with what safeguards.

Where it sits in the category. The other governance platforms in this appendix split roughly into runtime controls (TruthSystems, WitnessAI) and inventory or assessment tooling (FairNow, Holistic AI). CounselGuard occupies the law-firm-specific assurance and audit position: less focused on real-time prompt blocking, more focused on producing the defensible record across jurisdictional rule sets that bar counsel and clients will ask for. The vendor explicitly contrasts itself with generic GRC tools, emphasizing legal-jurisdiction rule mapping (ABA, SRA, FLSC), an AI tool registry built around the products law firms actually use, and bar-audit-ready evidence packages.

Practical evaluation questions for a legal buyer: how the jurisdiction-rule mapping is maintained as the underlying authority changes (ABA Opinion 512 will not be the last word; the EU AI Act timeline still has steps remaining; the Colorado picture is unsettled); how the tool registry stays current as new vendors and new model releases land; how integrations with the firm's identity, matter management, and document management systems work in practice; what the data-residency and confidentiality posture is for the audit-log content itself, since by design that content concentrates sensitive information about firm operations.

Verification and citation-checking tools

A separate category of tools focuses specifically on the verification problem discussed in Chapter 8. These include citation-verification add-ons that integrate with Westlaw, Lexis, or NetDocuments (the NetDocuments Legal Citations Verification App, LeanLaw and CiteTrue offerings, the LitigAI cite-checker, and the Lexis+ AI linked-citation feature). For firms whose dominant verification cost is citation checking, layering one of these tools on top of the drafting workflow is among the highest-ROI investments a program can make.

The right product depends on the dominant drafting platform and the practice mix. Litigation-heavy practices benefit most from deep Westlaw or Lexis integration; transactional practices benefit more from contract-focused tools (Spellbook, Harvey, CoCounsel) that have verification built into the workflow rather than bolted on.

Specialty legal AI platforms

The major specialty legal AI vendors (Thomson Reuters CoCounsel, Harvey, Lexis+ AI, Paxton, Spellbook, Legora, DeepJudge, Norm AI, GC AI, and others) increasingly compete on governance posture in addition to capability. CoCounsel publicly describes a multi-layered governance framework including data boundaries (no training of third-party foundation models on customer inputs), expert validation (4,500-plus subject-matter contributors), workflow integration in existing platforms like Westlaw and Microsoft 365, and a multi-model architecture that routes across frontier models from Anthropic, OpenAI, and Google plus proprietary models[30]. Harvey and Legora have made similar disclosures, and the recent funding rounds in this segment (Legora's reported $5.6 billion valuation in early 2026) suggest the category will continue consolidating around vendors that take governance seriously as a competitive differentiator.

How to think about buying any of these

Decide the operating model first. The platform should fit the program; the program should not be reshaped to fit the platform.

Insist on clear scope: inventory only, inventory plus assessments, inventory plus assessments plus runtime controls. These are different purchases, often confused in marketing.

Pressure-test regulation-mapping claims. Ask to see EU AI Act, NIST AI RMF, and applicable state bar guidance reflected in the product, not in the marketing.

Treat platforms as accelerators, not substitutes. A poorly governed program produces poorly governed data inside a beautifully designed platform.

Appendix B. Glossary

Plain-English definitions of the terms and acronyms used in this primer. Use it as a quick reference while you read, or hand it to a colleague who is new to the topic.

Agentic AI. An AI system that takes multi-step actions across tools or systems with limited per-step human approval. Different governance tier from response-only AI; covered in Section 13.

AI (artificial intelligence). Software that performs tasks typically associated with human reasoning: reading, writing, translating, summarizing, recommending. In this primer, references to AI mean generative AI unless stated otherwise.

AI RMF (NIST AI Risk Management Framework). The voluntary framework published by the U.S. National Institute of Standards and Technology that organizes AI risk work into four functions: Govern, Map, Measure, and Manage. The primer's recommended anchor framework.

Calibration. The match between a model's apparent confidence and its actual reliability. A well-calibrated model is uncertain when it should be uncertain. Most current frontier models are poorly calibrated by default; the lawyer's job is to do the calibration the model does not.

Change champion. A respected lawyer or staff member who carries the governance message into their practice group or business unit. Champions translate policy into daily practice and surface real friction back to the committee. Named champions are a quiet predictor of programs that actually land.

Change management. The discipline of moving an organization from one way of working to another on purpose, with communication, training, feedback loops, and deliberate sequencing. The primer treats AI governance as a change management problem before it is a documents problem.

Colorado AI Act. Colorado SB24-205, as amended by SB25B-004, the first U.S. state law placing affirmative duties on developers and deployers of high-risk AI systems. Effective June 30, 2026 in its current form.

Confabulation. The term NIST uses for what most lawyers call a hallucination: content that sounds authoritative but is not true. The classic example is a fabricated citation.

Culture (as used in this primer). The norms and behaviors that decide whether people actually follow the written policy when no one is watching. Treated alongside structure and change management as one of the three pieces of working governance.

DPIA (Data Protection Impact Assessment). A structured review of how a tool handles personal data. Required under the GDPR for high-risk processing and useful as a template even where the law does not require one.

Endpoint. The specific version of a model or service you actually call, including the hosting arrangement and configuration (training opt-out, logging policy, region).

Enterprise tier. A vendor's business version of a tool, typically sold under a negotiated contract. Offers training opt-out, administrative controls, data residency choices, and confidentiality commitments that consumer versions do not. Treated in this primer as the floor for any tool that touches client content.

Epistemic humility. Recognition that the lawyer also has limits in knowing when the model is wrong. A model can produce a plausible answer to a question whose correct answer requires expertise the lawyer lacks. The discipline is to refuse to use AI output for anything you cannot verify, even when the model sounds right.

EU AI Act. Regulation (EU) 2024/1689, the European Union's comprehensive AI law. Uses a four-tier risk structure (unacceptable, high, limited, minimal) plus a separate regime for general-purpose AI. Key dates: February 2025, August 2025, August 2026.

Fine-tuning. Training an existing model further on your own data to change how it responds. Creates real privilege and confidentiality questions because your data becomes part of the model.

GAI or GenAI (generative AI). AI that produces new content (text, code, images, audio) rather than only classifying or predicting. Most of the risks in this primer come from generative AI.

GPAI (general-purpose AI). Models that are not built for a single task and can be used across many applications. The EU AI Act has a distinct regime for GPAI models.

Hallucination. Common term for AI output that reads as confident fact but is invented. NIST calls this confabulation. Fabricated case citations are the best-known legal example.

HITL (human in the loop). A workflow that requires a qualified human to review and approve AI output before it is used or sent. Treated as mandatory for client-facing work product and any consequential decision.

Incident. An event that triggers your response plan. For AI purposes: a fabricated authority that left the firm, confidential content sent to a non-approved tool, a vendor breach, or a material model behavior change affecting a live use.

ISO/IEC 42001. The first certifiable international management system standard for AI. Plan-Do-Check-Act structure familiar from ISO 27001. An independent registrar can audit an organization against it.

Model inventory. The living register of every AI tool in use across the firm or department, including tools adopted informally by individual lawyers. The first document a client or regulator typically asks to see.

Model Rule. A rule from the ABA Model Rules of Professional Conduct, the template most U.S. states adapt. The six rules that matter most for AI are 1.1, 1.6, 1.4, 3.3, 5.1 and 5.3, and 1.5.

Multi-model architecture. A product design that routes requests across two or more underlying models (often frontier models from different vendors) depending on the task. Discussed in Section 10 in the context of orchestration platforms like CoCounsel.

OCG (outside counsel guidelines). A client's written rules for how outside counsel must run their matter, covering staffing, billing, conflicts, confidentiality, and increasingly AI.

Opinion 512. ABA Formal Opinion 512 (July 2024), the first formal ABA opinion on generative AI. Applies the existing Model Rules to GAI rather than creating new duties.

Output. What the model produces in response to a prompt. Under most legal frameworks, the lawyer owns responsibility for the output regardless of how it was generated.

Privilege (attorney-client). The protection that lets clients speak candidly with their lawyers without those communications being compelled in litigation. Disclosing privileged content to a third party (including an AI vendor without adequate protections) can waive it.

Prompt. The input you give the model. May contain instructions, context, and data. Treat prompts as potential exhibits: they may be logged by the vendor and may be discoverable.

RACI. A simple chart showing who is Responsible (does the work), Accountable (owns the outcome), Consulted (gives input), and Informed (needs to know). Section 4 includes a RACI for AI governance roles.

Reliability. The probability that a model will produce a correct output for the kind of task you are giving it, with the kind of inputs you are providing. Task-specific. A model highly reliable at summarization may be much less reliable at jurisdictional analysis.

Retrieval-augmented generation (RAG). A technique that pulls approved source documents at query time and supplies them to the model so the model can cite real material. Reduces hallucination risk; does not eliminate it.

Risk tier. The label assigned to a use case (low, elevated, or high in this primer) that determines what controls apply. Tiering stops low-stakes uses from being over-governed and high-stakes uses from slipping through.

Shadow AI. The use of unsanctioned AI tools by lawyers and staff inside the firm. Most often a response to the approved tool being unavailable or worse than the consumer alternative. Treated in Section 11 as a market signal rather than only a compliance failure.

Skill atrophy. The gradual loss of foundational skills (in this primer, foundational drafting and research skills in junior lawyers) when AI assumes the work historically used to develop those skills. Section 11 covers the development implications.

SOC 2 Type 2. An independent attestation that a vendor has operated its security controls effectively over a period (usually 6 to 12 months). The minimum security baseline this primer recommends for any AI tool touching client or confidential content.

Subprocessor. A third party a vendor uses to help deliver its service, such as a cloud provider or a model host. Subprocessor lists and change notices belong in any enterprise AI contract.

Training data. The content used to build or refine a model. If your data is used as training data, it can influence the model's future behavior for other customers. Most enterprise AI contracts should prohibit this.

Verification. The step where a qualified human confirms that citations, quotations, factual claims, and legal reasoning in AI-assisted work product are accurate. Opinion 512 and every state bar opinion this primer cites treat verification as non-negotiable. Section 8 decomposes verification into citation, quotation, factual, and reasoning checks.

Appendix C. Endnotes

Numbered references correspond to the superscript markers used in the body text. Figures and survey statistics are preserved with their reporting organization's name so readers can verify against the original source.

1. Association of Corporate Counsel and Everlaw, 2025 GenAI Survey, reporting active generative-AI use at 52 percent of corporate law departments (up from 23 percent in 2024) and that 64 percent of in-house teams expected to rely less on outside counsel as internal AI capability grew. Based on 657 in-house respondents across 30 countries. Press release.

2. Clio, Legal Trends Report and AI Policy resources, reporting majority adoption across the profession alongside roughly 44 percent of firms operating without a formal AI policy. Clio resource hub.

3. FTI Technology, 2025 General Counsel Report, finding only 15 percent of GCs surveyed feel prepared to manage AI risk. FTI Technology.

4. Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act); see European Commission AI Act Service Desk implementation timeline. Prohibited practices apply Feb. 2, 2025; general-purpose AI obligations Aug. 2, 2025; most remaining provisions and Commission enforcement powers Aug. 2, 2026. EC implementation timeline.

5. Colorado SB24-205 (Colorado AI Act), as amended by SB25B-004, deferring the effective date from Feb. 1, 2026 to June 30, 2026. Colorado General Assembly.

6. Mayer Brown, The Colorado AI Policy Work Group Proposes an Updated Framework to Replace the Colorado AI Act (March 2026). Mayer Brown.

7. ABA Standing Committee on Ethics and Professional Responsibility, Formal Opinion 512, Generative Artificial Intelligence Tools (July 29, 2024). Applies Model Rules 1.1, 1.6, 1.4, 3.3, 5.1/5.3, and 1.5 to generative AI use. ABA PDF.

8. California State Bar, Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law (Nov. 2023). Cal Bar PDF.

9. Florida Bar Op. 24-1 (Jan. 19, 2024), Florida Bar; NYC Bar Formal Op. 2024-5, NYC Bar; North Carolina State Bar 2024 Formal Ethics Opinion 1; Oregon State Bar Formal Op. 2025-205.

10. Justia, AI and Attorney Ethics Rules: 50-State Survey, Justia; NYC Bar, Analysis of Current Ethics Guidance Related to Generative AI (May 2025), NYC Bar PDF.

11. UK Solicitors Regulation Authority, repeated guidance on generative AI in legal practice; Bar Council of England and Wales guidance for barristers; opinions of EU national bars (France, Germany, the Netherlands) tracking the ABA position with GDPR-specific cautions. See ABA Task Force on Law and Artificial Intelligence, Year Two Report (2025), available via americanbar.org.

12. Standing Order of Judge Brantley Starr, U.S. District Court for the Northern District of Texas, Mandatory Certification Regarding Generative Artificial Intelligence (May 30, 2023). See Bloomberg Law tracker.

13. Standing Order of Judge Michael M. Baylson, U.S. District Court for the Eastern District of Pennsylvania, Regarding the Use of Generative Artificial Intelligence (June 2023).

14. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023). $5,000 sanction for fabricated ChatGPT citations.

15. Park v. Kim, 91 F.4th 610 (2d Cir. 2024). Discipline referral after similar conduct.

16. Public trackers of AI hallucination and sanctions cases: Damien Charlotin AI Hallucination Cases Database; Law360 Pulse AI Tracker; ABA Litigation Section, AI Hallucinations Are Real and How to Avoid Them, ABA. Aggregate counts vary by tracker and definition; cite the underlying opinions rather than aggregate figures.

17. NIST, AI Risk Management Framework 1.0 (NIST AI 100-1, Jan. 26, 2023). NIST PDF.

18. NIST, Artificial Intelligence Risk Management Framework: Generative AI Profile (NIST AI 600-1, July 26, 2024). NIST PDF.

19. ISO/IEC 42001:2023, Information technology - Artificial intelligence - Management system (December 2023). ISO.

20. Magesh, Surani, Dahl, Suzgun, Manning, Ho, Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools (Stanford, 2024); see also Stanford HAI summary, AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries. Stanford PDF; Stanford HAI.

21. Survey gap between actual AI use and firm-provided tooling synthesized from Embroker year-end 2024 reporting (showing 78 percent of US firms not yet adopting any AI tools while individual lawyer use rose); Relativity Blog, What Legal Leaders Should Know About Shadow AI; OneAdvanced, How Shadow AI Affects Law Firm Compliance. Relativity; OneAdvanced.

22. Gartner, General Counsel Should Assess AI Insurance to Mitigate AI Risks (April 2026). Gartner.

23. TruthSystems, Unified AI security and governance for legal practice. truthsystems.ai; coverage in Law.com, Legal Tech Startup Truth Systems Announces Tool for Preventing AI Misuse in Law Firms (Aug. 2025).

24. TruthSystems, company announcement: TruthSystems Raises $58 Million for Global Expansion and Announces New Ways to Secure AI Agents (January 2026). Verify scope and controls before procurement.

25. WitnessAI, Unified AI Security and Governance Platform. witness.ai.

26. FairNow, AI Governance Platform. fairnow.ai.

27. UK Government, AI Assurance Techniques: FairNow AI Governance Platform listing. gov.uk.

28. Holistic AI, enterprise AI governance platform. holisticai.com.

29. CounselGuard, AI governance and compliance platform for law firms. counselguard.io. Modules described include Dashboard, AI Tools, Compliance, Policies, Training, Activity, Reports, and Audit Log; jurisdiction mapping covers ABA Opinion 512, EU AI Act, SRA Standards (UK), FLSC Model Code (Canada), and Australian conduct rules. Product is in early-access onboarding via waitlist.

30. Thomson Reuters, CoCounsel governance and multi-layered framework. Thomson Reuters product page; see also Thomson Reuters, One Million Professionals Turn to CoCounsel as Thomson Reuters Scales AI for Regulated Industries (February 2026), press release.