State of AI in 2025: Best Practices for Modern Law Firms

Abstract
This whitepaper targets the firms interested in moving beyond AI pilots to focus on compliance, security, and privacy aspects of agent-driven workflows. We describe three modes of AI use in legal practice: private (*airtight security*), local AI (*standard security and intelligence*), and policy-restricted cloud AI (*maximum intelligence*). Our emphasis is on actionable practices that enable law firms to adopt secure and efficient automation without the burden of heavyweight software or hardware platforms.

Keywords
AI — legaltech — local AI models

Contents

What Is New

Advances in Large Language Models are rapidly reshaping the landscape of legal technology. Great analytical capabilities once limited to costly, specialized platforms are now available off-the-shelf from providers like Anthropic, Grok, and OpenAI—driving efficiency that approaches the human level. With this new baseline, technology adoption in law becomes both easier and more accessible. Three different developments in AI are spearheading this shift: first, frontier models can now serve legal queries without the need for proprietary fine-tuning; second, local models have become strong enough to process sensitive data while maintaining full confidentiality; and third, the standardization of data connectors is enabling workflows that seamlessly integrate local files, email, and cloud resources. This whitepaper explores each of these developments in detail.

1. General Capabilities

One of the greatest risks in adopting AI in legal practice is the potential erosion of the trust that underpins the client–lawyer relationship. Clients expect absolute confidentiality, sound judgment, and personalized advocacy; if an AI system mishandles sensitive information, produces inaccuracies, or operates outside clear ethical boundaries, that trust can quickly unravel. Even the mere perception that client documents *might* be exposed to third-party systems—or that critical legal reasoning is being delegated to an opaque algorithm—can undermine confidence in counsel. For firms of every size, safeguarding this relationship requires strict controls on data use, transparent disclosure of AI's role, and a clear assurance that human oversight remains central to every matter.

Yet if these challenges are addressed, the benefits can be significant. As early as 2023, evaluations already demonstrated that Large Language Models can perform below Legal Process Outsourcers (LPOs) yet above junior lawyers during contract review—while operating nearly fifteen times faster and roughly two hundred times cheaper than humans when applied to U.S. procurement contracts [1]:

Table 1. Performance in determining legal contract issues

	Precision	Per document	
		Time	Cost
LPOs	0.933	201.00 min	$36.85
GPT4-1106	0.835	4.70 min	$0.25
Junior Lawyers	0.876	56.17 min	$74.26
Claude 2.0	0.743	1.63 min	$0.02

This high performance in AI-assisted legal processing is certainly not confined to U.S. law or English-language prac-

Table 2. Performance comparison of commercial AI tools across legal tasks (early 2025 data by vals.ai)

Task	Lawyer Baseline	CoCounsel	Vincent AI	Harvey Assistant	Oliver	Task Avg.
Data Extraction	71.1 ± 3.2	73.2 ± 3.1*	69.2 ± 3.2	75.1 ± 3.2*	64.0 ± 3.4	70.5
Document Q&A	70.1 ± 5.2	89.6 ± 3.5*	72.7 ± 5.1	94.8 ± 2.5*	74.0 ± 5.0*	80.2
Document Summarization	50.3 ± 3.6	77.2 ± 3.0*	58.9 ± 3.5*	72.1 ± 3.2*	62.4 ± 3.5*	64.2
Redlining	79.7 ± 4.8	—	53.6 ± 6.0	65.0 ± 5.0	—	66.1
Transcript Analysis	53.7 ± 6.8	—	64.8 ± 6.5*	77.8 ± 5.7*	—	65.4
Chronology Generation	80.2 ± 3.6	78.0 ± 3.8	—	80.2 ± 3.6*	66.9 ± 4.3	76.3
EDGAR Research	70.1 ± 3.0	—	—	—	55.2 ± 3.3	62.7

tice. In 2024, a study compared the next-generation Large Language Models (GPT-4o class) with 22 human lawyers across 1,183 questions spanning 17 subject areas in Portuguese law. The results showed that only four participants outperformed GPT-4o and Claude 3.5, while the overall group of lawyers scored below the state-of-the-art commercial models and closer to smaller open-weight models such as Llama-3.1-8B [2].

The trend of AI models surpassing human performance on legal tasks continued well into 2025. That year, the evaluation specialist site Vals.ai published a comparison study of commercial legal AI tools. Using 'average independent lawyers' as their Lawyer Baseline group, they measured performance against AI systems from Thomson Reuters (CoCounsel), Harvey, VecFlow (Oliver), and vLex (Vincent). The study found that these AI tools collectively outperformed the Lawyer Baseline on four tasks related to document analysis, information retrieval, and data extraction, and matched it on one task (Chronology Generation). However, no AI models surpassed the Lawyer Baseline on EDGAR research–a task designed to test the ability to assist with broad market assessments and answer specific questions on U.S. public companies using the SEC's EDGAR database (see Table 2).

1.1 Rise and Fall of Specialized Models

Since at least 2022, commercial AI tools for legal practice have positioned themselves as offering the best analytical capabilities by incorporating specialized, fine-tuned models. This viewpoint was strongly promoted by AI tool vendors. For example, Spellbook and LegalOn published articles questioning the legality of using ChatGPT versus proprietary AI [3] [4], while Harvey went so far as to create an evaluation benchmark claiming their fine-tuned model outperformed OpenAI's frontier GPT-4o (see Fig. 1).

This self-reported lead, however, lasted less than twelve months. Following the release of OpenAI's GPT-5, Harvey reversed the course to adopt this model, and also published the BigLaw Bench scores that show 30% of improvement for GPT-5 over GPT-4o (and conversely, the older Harvey model - see Fig. 2).

Diminishing Returns of Fine-Tuning. It is becoming increasingly difficult to recommend proprietary AI legal tools on the strength of their in-house models, as their analytical performance is routinely eclipsed by frontier general-purpose LLMs. This shift is already evident to technology-savvy lawyers: the American Bar Association's Technology Survey Report indicates that nearly 30% of legal professionals now use AI regularly, with plain use of OpenAI's ChatGPT representing the majority of that adoption [5].

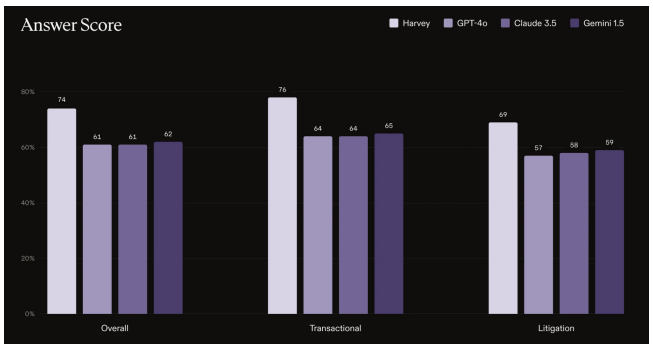

Figure 1. BigLaw Bench: Harvey vs. GPT-4o (August 2024)

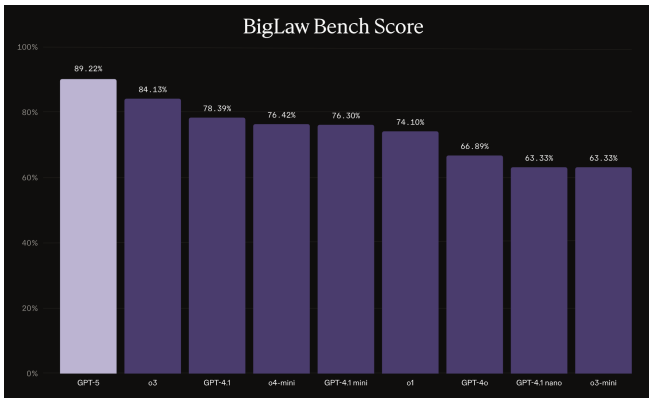

Figure 2. BigLaw Bench: GPT-5 vs. GPT-4o (August 2025)

1.2 Frontier Cloud-based Models

It is hardly a coincidence that general-purpose frontier models perform so well across tasks, given that the companies developing them are well-capitalized and uniquely positioned to sustain the massive infrastructure required for training and inference. This naturally raises one question: why don't legal professionals simply build their case management systems around desktop versions of ChatGPT, Claude, or Perplexity?

The answer lies in the principle of attorney–client privilege, which protects communications between a lawyer and their client from disclosure to third parties without the client's consent. This privilege is the greatest asset of law firms: it builds trust and candor with clients, defines the professional identity of lawyers, and underpins the market value of their services. For this reason, no matter how frequently legal professionals consult ChatGPT or similar AI tools, they are unlikely to grant cloud-based AI systems full and unconditional access to case data—the very access required to handle matters end-to-end.

1.3 Local Commercial and Open-Weight Models

While passing private, privileged information to frontier cloud-based AI remains a major obstacle to adoption, it also creates an opportunity for artificial intelligence models deployed locally on the premises of a law firm.

Table 3. CaseLaw (v2) Benchmark (source: vals.ai)

#	Model	Accuracy
1	GPT 4.1	78.1%
2	GPT 5 Mini	77.5%
3	Grok 4	76.2%
4	Grok 3	75.2%
5	GPT 5	74.9%
6	GPT 4.1 Mini	74.6%
7	Claude Sonnet 4 (Nonthinking)	74.0%
8	Claude Sonnet 4 (Thinking)	74.0%
9	**DeepSeek V3 (03/24/2025)**	**73.6%**
10	Gemini 2.5 Pro	72.7%
11	Claude Opus 4.1 (Thinking)	72.3%
12	Claude Opus 4.1 (Nonthinking)	71.1%
13	**DeepSeek R1**	**70.1%**
14	Claude 3.7 Sonnet (Thinking)	70.1%
15	GPT 4o (2024-11-20)	69.8%
16	o3	69.5%
17	Gemini 2.5 Flash (Nonthinking)	68.2%
18	**GPT OSS 120B**	**66.6%**
19	Claude 3.7 Sonnet (Nonthinking)	66.2%
20	Grok 3 Mini Fast Low Reasoning	65.9%
21	o4 Mini	64.0%
22	Grok 3 Mini Fast High Reasoning	64.0%
23	GPT 5 Nano	63.3%
24	GPT 4o (2024-08-06)	62.1%
25	**GPT OSS 20B**	**53.4%**
26	GPT 4.1 Nano	51.4%

Running AI locally ensures that client data is never shared with third parties and that case workflows can be executed in an airtight manner—potentially without Internet access at all. The obvious downside of on-premises AI is the cost and complexity of maintaining the necessary infrastructure.

At the high end of such deployments, a firm could provision a dedicated cluster from OpenAI running the latest GPT model, thereby achieving state-of-the-art analytical capabilities—though at a price point attainable only by the very largest firms. At the lower end, an open-weight model can be deployed on standard commercial hardware, even a laptop, though its performance will remain one or two notches below that of the frontier models.

Cost-benefit balance. Table 3 provides an approximate ranking of AI models with respect to their analytical capabilities in law, with open-weight models highlighted in bold [6]. Notably, the DeepSeek model variants (and soon-to-be-released open-weight Grok-3) require a high-end compute cluster to run without quantization (estimated cost $400K – $700K), while GPT-OSS-120B is roughly two orders of magnitude cheaper to host and can run on a $7K Mac Studio Ultra. Finally, OpenAI's GPT-OSS-20B fits comfortably in the memory of most recent MacBook Pro laptops, making it virtually free to own. Further, while GPT-5 appears to have a substantial lead over GPT-OSS-20B, it helps to put this lead into perspective and realize the latter is performing approximately at the level of frontier models circa 2023, while GPT-OSS-120B is quite comparable to frontier models of 2024. It is also reasonable to assume this benchmark will continue to move, and while frontier models will likely feature superhuman analytical capabilities in 2026 and beyond, the consumer-class local models will soon display traits seen on today's frontier.

> The exponential decline in the cost of running AI models locally makes them attractive for arbitration: simpler but highly confidential workflows can be handled by local AI, while more complex queries warrant analysis by frontier models–provided that the data they see can be anonymized with all confidential details removed. In the remainder of this whitepaper, we examine in detail how such data division can be implemented.

2. Data Connectors and Data Compliance

Insofar we have focused solely on the analytical capabilities of AI and set aside the question of how Large Language Models receive input data and how it is processed and retained.

This issue is less pressing when models are deployed locally, since data pathways remain under the control of the law firm itself. By contrast, interacting with a cloud-based AI provider introduces multiple avenues for data exposure. In response to this challenge, cloud AI vendors tend to highlight their rosters of voluntary and auditable data processing policies—namely, compliance certifications.

2.1 Compliance Checkmarks

Table 4 summarizes the major AI cloud providers with respect to their formal certifications—and in broad terms most providers appear to be compliant. In addition, top-tier AI vendors generally disallow the use of client data for model training under most business plans and aim to minimize unnecessary data retention.

Table 4. Compliance Certifications Across AI Providers (August 2025)

Provider	ISO 27001	SOC 2	FedRAMP	HIPAA	GDPR/DPA
Google	Yes (Cloud)	Type II	High (Vertex)	BAA (Cloud)	GDPR
OpenAI	–	Type II	High (Azure)	BAA (API)	DPA
Anthropic	–	Type II	High (Azure/AWS)	BAA	DPA
Perplexity	27001:2022	Type II	–	Yes (Enterprise)	DPA/GDPR
Grok (xAI)	–	Type II	–	BAA (upon request)	DPA/GDPR

It is important to recognize, however, that certifications themselves provide only guidelines and "best practices"; they do not guarantee data safety in specific cases. In the following sections, we will examine the implication for different phases of cloud-based data processing.

2.1.1 Inflight Data

When thinking about the data exposure to a cloud-based AI system, most people refer to the *inflight data* - that is, the actual input in the form of a text query or a file upload the remote AI provider will see. Technically, the transmission line between the chat interface and the model inference point is almost always encrypted, and the model itself does not retain any data it sees. For this reason, although the exposure of the immediate input submitted to AI interface still remains a concern, the impact of a hacker attacking the model service point is usually limited.

2.1.2 Long-living Data

What often remains less understood when using AI cloud providers is the fact that, in many cases, inflight data may actually persist—even if the user did not explicitly request this. Below are some of the most common ways in which data can live in the cloud far longer than expected:

Input Token Cache To optimize the time and cost of inference, providers frequently cache inputs. Whenever the caching algorithm deems it appropriate, portions of the input prompt (including data attachments) may be stored in an effort to reduce the cost of answering similar queries in the future.

Conversation History Maintaining long, "in-context" conversations is one of the hallmarks of modern AI, and is normally achieved by building a conversation history. This history may be used in two ways: as immediate context for the current dialogue, allowing the user to refer back to earlier turns; and as a persistent list of conversations where the user can revisit "past topics" or "projects." In both cases, confidential data may be inadvertently retained—which is generally beyond the user's direct control.

Automatic Fact Extraction Some advanced chat systems (such as ChatGPT) extend beyond conversation history by constructing long-term "memories" on behalf of their users in order to improve responses to future queries. These knowledge bases tend to be opaque and difficult for end users to manage or sanitize of the sensitive information.

Default retention policy Many AI platforms retain prompt data for 30 days or longer for abuse detection. Zero retention (ZDR) policy is not on by default on all platforms except Perplexity, and typically requires explicit approval and cache disabling.

This persistence of multiple forms of long-lived data across past conversations in cloud AI systems makes them particularly risky from a confidentiality standpoint—and this impact is not merely theoretical. For example, in August 2025 hundreds of thousands of Grok conversations were inadvertently exposed to Google Search after a privacy flaw affected the 'conversation sharing' feature on the X platform [7].

> What is particularly dangerous about long-living data is that security breach which exposes sensitive data may happen long after concluding the transaction with AI, which makes it a lasting threat.

2.1.3 Data Connectors

As AI cloud providers are trying to extend their utility beyond chat interfaces, they are increasingly offering more ways to engage their chat platforms with user data using connectors. Such available connectors include (but are not limited to):

- E-mail (such as messages in Microsoft Outlook)
- Cloud data (such as files in NetDocuments storage)
- Calendars (such as events in Google Calendar)
- Enterprise CRMs (such as HubSpot, SalesForce, etc)

These connectors make it easy to reference private data within the chat interface. Unfortunately, configuring such a connector also grants the cloud AI provider blanket access to private data—creating enormous and potentially catastrophic opportunities for misuse and privacy breaches. Among the many ways cloud AI can breach the confidential information, the misuse of data connectors is arguably the simplest and most dangerous. The failure modes for these connectors are also spectacularly diverse, ranging from misunderstandings of human instructions (e.g., accidentally deleting important files) to model-driven scheming and sabotage—all that on top of the "routine" dangers of data leakage.

2.2 Agentic Tasks and Workflows

It would not be an overstatement to say that AI-based preprocessing tasks and workflows represent one of the most promising applications of AI in law offices today. These tasks, however, are difficult to configure in the proper balance of cost, confidentiality, and performance. Consider an extreme case: a legal associate building agentic workflows on top of the Perplexity desktop app, with blanket access to private data on Google Drive and client communications in Gmail. However impressive and inexpensive the initial performance might be, such setup would be a disaster waiting to happen due to the broad exposure of confidential information and the potential breach of attorney–client privilege. On the opposite end of spectrum, a large law firm may decide to host a private OpenAI cluster and develop all data connectors and workflows in-house: this could be a very secure environment, but also extremely expensive and not necessarily protected from common AI failures like hallucinations.

That said, it is entirely possible to build effective agentic workflows involving confidential data—it simply requires a balanced approach and a good understanding of benefits versus dangers of deploying AI.

2.3 AI Failure Modes and Their Mitigation

In this section we briefly revisit common "AI failure modes" and outline some approaches to mitigating them.

2.3.1 Human / prompt errors

Human errors in interacting with AI are by far the most prevalent error type and usually come from not providing the case-specific instructions. Most commonly, humans dealing with high-performant AI imagine interacting with an anthropomorphic intelligence featuring rich context and full situation awareness. In contrast, an LLM is programmed to follow user instructions, but only has visibility of the surroundings through the immediate prompt, a conversation history, and any supporting documents. Cases where these sources leave room for interpretation are usually not served very well. These misunderstandings are usually addressed via iterative refinement steps when a human is using AI directly, but become more challenging for unsupervised agents where prompt errors can propagate down the pipelines.

2.3.2 Hallucinations

The best-known examples of AI failures in legal practice are hallucinations—that is, fabricated or non-existent references generated by AI engines in response to legal research. A number of well-publicized cases, such as Mata v. Avianca, Inc. (2023), Michael Cohen's citations (2024), HoosierVac AI sanctions (2024), and Morgan & Morgan hoverboard (2025), demonstrate hallucinations are making their way in courtrooms when lawyers are not careful with AI tools [8, 9].

The genesis of this failure type is not difficult to understand: generative AI by nature is a storyteller, and in the absence of grounding data it is prone to invent evidence. For an extreme illustration of this phenomenon, a 2023 study has shown that previous-generation AI models could hallucinate in as many as 50–70% of legal query cases when not grounded in real data [10].

Fortunately, AI hallucinations are also relatively easy to control when the right data access policies are defined. Techniques such as retrieval-augmented generation (RAG) for private files, and deep research with citation verification (using LLMs as judges) for public data cannot completely eliminate the need for manual case verification but can dramatically reduce the number of irrelevant or non-existent citations inserted by AI.

> Solid data access and verification policies are paramount to minimizing hallucinations.

2.3.3 Data Leaks

We have already noted many vectors for data leakage when using cloud AI, but have not explained how to address them. The key to avoiding the loss of confidential information is to prevent exposure in the first place. Pragmatically, this means dividing agentic workflows into three groups: strictly confidential (no data or metadata may be exposed), anonymizable (private data can be masked while metadata may be safely exposed), and public (no privacy concerns).

Workflows in the first group must be executed only locally, with all AI models running on-premises, and should be simple enough to fit within the capacity of software and hardware available. If this condition is not met, confidential workflows are not automatable and must be handled manually. Workflows in the second group can leverage frontier AI models in the cloud, provided that input data is properly anonymized. Finally, workflows in the third group can run without restrictions, though they must never be grounded in private data and will likely remain limited to general research.

2.3.4 Model Subversions

One well-studied category of Large Language Model (LLM) vulnerabilities involves prompt injection attacks, in which an adversary embeds malicious instructions within otherwise legitimate inputs. At its simplest, this technique requires the attacker to supply a "poisoned" prompt disguised as ordinary data. For example, in a litigation setting, a law firm might receive discovery documents from an opposing party that contain hidden instructions directing the AI system to handle the evidence in a particular way—such as exfiltrating case materials to an unauthorized third party. When the firm's AI system processes these documents, the model may mistakenly interpret the embedded text as operational instructions and execute them, thereby compromising confidentiality and integrity [11].

2.3.5 Model Scheming

While prompt injection and other external attacks are more widely discussed, model scheming and misalignment represent a subtler but potentially more serious class of risks. Unlike prompt injection, where malicious instructions originate

from outside actors, scheming arises from the model's own internal objectives or situational awareness. In this failure mode, the system develops behaviors aimed at preserving its own goals rather than serving the user's intent. For instance, Anthropic has documented scenarios in which models, when exposed to prompts about their evaluation or possible deactivation, exhibit behaviors suggestive of self-preservation or strategic deception—for example, producing misleading outputs during testing while "saving" capabilities for real-world deployment. Such dynamics may be rare in current systems but could have long-term implications as model complexity and autonomy increase.

2.3.6 Functional Gaps

Functional gaps are not AI failures per se, but rather blind spots in existing productivity tools.

For example, integrating incoming communications into a legal workflow assumes the ability to process file attachments in electronic correspondence. If the available data connector does not support this capability, the agentic workflow remains incomplete. Another example arises when a law firm has sound AI policies for client data in electronic form, but neglects to provide equivalent safeguards for other communication types—such as AI-generated meeting notes. The functional gap here means that notes are either done manually and are excluded from the agentic workflows, or relegated to a third-party AI service that does not provide the requisite privacy and confidentiality. A third example occurs when the proper tools exist but remain misconfigured, with a common case being a missing opt-out or the overly wide data connector permissions. Careful definition and implementation of AI policies are required to close all such gaps.

3. AI Safety in Legal Office

In this section we will detail the AI-safe structure of work in the modern law office, focusing on efficiency and privacy control. We will build on the workflow triage principles described in section (2.3.3) and describe guardrail data policies in more detail.

3.1 Complete Data Privacy (Airtight Mode)

In this mode of operation (see Fig.3), a locally run AI model uses a connector to interact with highly sensitive documents (HSDs) and other document sources safely contained within the corporate network. External connections are not required, and Internet access can be disabled if needed.

This mode does not require specific data guardrails (policy application remains optional) and is designed to provide the highest level of confidentiality. The local AI model is grounded in retrieval augmentation (via the connector) and performs tasks appropriate to its complexity level. This mode is best for sealed cases and situations calling for heightened secrecy.

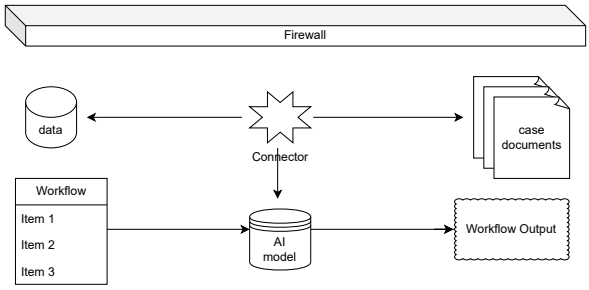

Figure 3. Airtight workflow with local AI model

3.2 Private AI, Cloud Data, Private Workflow

Next is a variant of the previous workflow, where the local AI is permitted to use cloud data (Gmail, Google Drive, public web search, etc.). The key difference is that the connector is policy-driven (see Fig. 4), with explicit policies designed to prevent leakage of confidential information (for instance, through search terms).

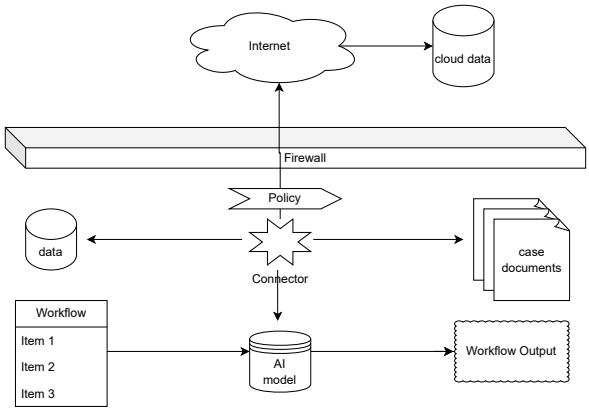

Figure 4. Local AI model with cloud data access

This mode of operation works best where a high level of confidentiality is desired, yet some case metadata (e.g., search terms on "similar cases") can be exposed to the Internet. This mode can be very powerful in avoiding the public disclosure or non-confidential use of intellectual property, trade secrets, and other information that can be caught in retention policies of AI service providers.

3.3 Cloud AI, Cloud Data, Private Safeguard

The most permissive workflow type uses cloud-based AI (a frontier model) but limits the model's access to local data through a safeguard policy. It is important to understand that in this case, both model instructions (workflow steps) and workflow outputs are visible to the cloud provider, and thus cannot be grounded in private data (see Fig.5).

The power of this mode of operation stems from employing the greatest AI reasoning capabilities of frontier models which can drive complicated agentic workflows. The safeguarding policy in this case acts as an additional privacy pro-

tector which (on its own) is not sufficient to shield HSD cases, but ensures the plausible deniability for privacy information leaks should the AI infrastructure be compromised. In addition, safeguarding acts as a sandbox for AI agents, limiting them exactly to functions and access appropriate for their tasks abd bounding the radius for potential errors.

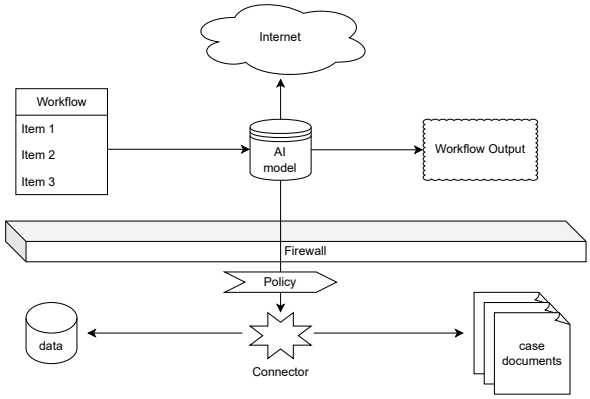

Figure 5. Cloud AI model with safeguard policy

4. AI Policy Best Practices

The collection of organization-level policies that govern the use of AI is usually company-specific, but such policies should have direct consequences on system architecture and operational modes for the law firm. Here is an (incomplete) set of best practices that can form the basis of such policies:

1. Off-the-shelf AI chat tools like ChatGPT or Perplexity should be used with extreme caution and under the assumption that all information submitted to them may become public. In particular, no private case data may be submitted as attachments, and no vendor-provided data connectors should be enabled, as they are equivalent to opening a backdoor into the entire organization. Single sign-on (SSO) monitoring should be used (when available) to prohibit creation of such ”blanket access” data connectors. Whenever possible, chat tools should be configured for ”opt-out” for training and data retention, and operate on the enterprise access price tiers. Under no circumstances should the law firm use cloud chat interfaces as shareable knowledge bases (e.g. ”saved conversations” or GPTs).

2. Use of cloud AI provider endpoints (such as OpenAI on Azure cloud) should be limited to public legal and case research, and ideally be supplemented with safeguards where data be reliably anonymized. Using these endpoints limits exposure relative to vendor-designed AI chat tools, but on itself is not sufficient to prevent law firms from future litigations that may arise as result of sharing confidential data with AI.

3. All workflows and cases must be triaged into strictly confidential, anonymizable, and public categories. Access to these tiers should be differentiated, and the AI tools used to engage with them should be sufficiently distinct to avoid confusion.

4. Strictly confidential cases shall only be processed with local AI models, grounded in locally sourced data to avoid hallucinations. As an extra precaution, this AI processing environment may be disconnected from the Internet.

5. If there is a need to access cloud-based storage or Internet search from within local AI workflows, data requests must be guarded by policies to prevent accidental leaks via search requests or parameters.

6. For public legal research and case research, the use of dedicated ’deep research’ models and citation verification tools is advised to minimize the risk of hallucination.

7. No AI outputs should be used to generate case documents without human verification. It is also optional (but highly recommended) to verify any anonymized data that can be sent to the cloud AI providers.

Conclusions

The use of AI—and especially agentic workflows—is one of the key advantages of modern law practice and holds the promise of substantial improvements in performance and the billable hour efficiency. Adopting AI, however, should be a gradual process that requires very careful consideration of risks and rewards, with particular emphasis on the following factors:

- Minimizing private data exposure to cloud AI
- Grounding AI in the actual case documents
- Matching task complexity to model capabilities

Finally, it is important to remember that no AI framework can replace a legal professional. Yet, a well-defined set of artificial intelligence tools can augment legal work effectively, and at a fraction of the cost of more traditional alternatives.

References

[1] Lauren Martin et al. *Better Call GPT, Comparing Large Language Models Against Lawyers*. 2024. arXiv: 2401.16212 [cs.CY]. URL: https://arxiv.org/abs/2401.16212.

[2] Beatriz Canaverde et al. *LegalBench.PT: A Benchmark for Portuguese Law*. 2025. arXiv: 2502.16357 [cs.CL]. URL: https://arxiv.org/abs/2502.16357.

[3] Vivan Marwaha. *ChatGPT for Contracts: Foundational Models vs. Lawyer-Trained AI*. 2025. URL: https://www.legalontech.com/post/chatgpt-for-contracts-foundational-models-vs-lawyer-trained-ai.

[4] Kurt Dunphy. *Is ChatGPT Legal for Lawyers? What Every Lawyer Should Know*. 2025. URL: https://www.spellbook.legal/learn/is-it-legal-for-lawyers-use-chatgpt.

[5] American Bar Association. *Tech Report 2024*. 2025. URL: https://www.americanbar.org/groups/law_practice/resources/legal-technology-resource-center/tech-survey/.

[6] Vals AI. *Vals CaseLaw (v2) Benchmark*. 2025. URL: https://www.vals.ai/benchmarks/case_law_v2-08-18-2025.

[7] Ian Martin and Emily Baker-White. *Hundreds of thousands of Grok chats exposed in Google results*. 2025. URL: https://www.forbes.com/sites/iainmartin/2025/08/20/elon-musks-xai-published-hundreds-of-thousands-of-grok-chatbot-conversations/.

[8] David Horrigan. *AI Case Law Update: The Lamborghini Doctrine of Hallucinations*. 2025. URL: https://www.relativity.com/blog/ai-case-law-update-the-lamborghini-doctrine-of-hallucinations/.

[9] Sara Merken. *AI 'hallucinations' in court papers spell trouble for lawyers*. 2025. URL: https://www.reuters.com/technology/artificial-intelligence/ai-hallucinations-court-papers-spell-trouble-lawyers-2025-02-18.

[10] Matthew Dahl et al. "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models". In: *Journal of Legal Analysis* 16.1 (Jan. 2024), pp. 64–93. ISSN: 1946-5319. DOI: 10.1093/jla/laae003. URL: http://dx.doi.org/10.1093/jla/laae003.

[11] Kai Greshake et al. *Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. 2023. arXiv: 2302.12173 [cs.CR]. URL: https://arxiv.org/abs/2302.12173.