The Seven Options Available to Attorneys to Use LLMs and Comply with Confidentiality Requirements

Introduction

Attorneys considering large language models (LLMs) face a central problem: how to use these tools without breaching confidentiality. Attorney–client privilege and professional responsibility rules make it essential to prevent client information from being disclosed, stored, or reused by technology vendors. At the same time, LLMs offer genuine advantages for research, drafting, summarization, and client communication.

This document sets out seven realistic options. It avoids speculation about exact costs or performance numbers because those vary dramatically by firm, region, vendor, and time. Instead, it highlights the structure of the choices, the potential pitfalls, and the main trade-offs. The aim is to equip lawyers to ask the right questions and choose a path that fits their practice.

1 Option 1. Zero Data Retention Agreements with Major Vendors

Description

Major providers such as OpenAI, Anthropic, Google, Microsoft, and Amazon Web Services offer contractual or technical options where data sent to the model is not stored or used to train future systems. These are called zero data retention (ZDR) arrangements.

Strengths

- Access to the most powerful models currently available.
- Clear contracts that specify retention and training policies.
- Lowest operational burden: you use the vendor's API or web interface.
- Costs scale with usage; no large upfront investment.

Weaknesses

- Requires trust in vendor compliance and internal practices.
- Settings must be correctly configured to ensure ZDR actually applies.
- Some logging for abuse prevention may still occur.
- Dependence on third-party vendors who may change policies.

Logistics

- Typically involves signing an enterprise contract or enabling a ZDR option in the administrative console.
- Attorneys must ensure all staff accounts follow the same settings.
- Data egress and audit logging should be reviewed with IT staff.

Pricing

- Usage-based, per million tokens of input and output.
- Prices vary but are generally in the low single-digit dollars per million input tokens and higher for output tokens.
- Volume discounts are available.

2 Option 2. On-Premises High-End Workstations

Description

A law firm can buy powerful desktop computers with advanced graphics cards and run open-source models locally. These machines cost roughly the same as a high-end gaming computer.

Strengths

- Complete control: no data leaves the building.
- Fixed, predictable cost after purchase.
- Suitable for experiments, drafting, and firm-internal work.

Weaknesses

- Limited power compared to cloud-grade hardware.
- Can only run smaller or heavily compressed models.
- Requires maintenance, patching, and security.
- Generates noise and heat.

Logistics

- Hardware purchase around five thousand dollars per machine.
- Electricity costs are modest but continuous usage adds up.
- IT staff must maintain systems and secure them against intrusion.

Pricing

- One-time hardware outlay plus electricity and staff time.
- No per-token costs, but capability is capped by hardware limits.

3 Option 3. Hosted Dedicated Machines

Description

Specialty providers and cloud vendors rent entire machines equipped with advanced GPUs. You have exclusive access to the machine but it is physically located in a data center.

Strengths

- More power than local consumer hardware.
- Quieter and physically secure off-site.
- Provider handles cooling, power, and network.
- Full control over software environment.

Weaknesses

- Requires technical skill to install and manage models.
- Still some dependence on provider infrastructure.
- Provider may be subject to subpoenas for logs.

Logistics

- Provisioned as bare-metal servers or dedicated virtual machines.
- Remote access via secure connection.
- Attorneys remain responsible for data handling inside the machine.

Pricing

- Charged hourly for GPUs; costs vary widely.
- Cheaper if you keep machines busy; expensive if idle.
- No per-token cost if you run your own model, but you pay for hardware time.

4 Option 4. Managed Services with Confidentiality Guarantees

Description

Services like AWS Bedrock, Google Vertex, and Cohere offer managed access to LLMs with commitments not to use customer data for training and to apply strong confidentiality controls.

Strengths

- Turnkey: little or no IT burden.
- Strong contractual protections.
- Access to a curated set of models optimized for business use.
- Integration with compliance frameworks.

Weaknesses

- Limited model selection compared to the open market.
- Higher prices than raw GPU rental.
- Must trust provider's implementation of confidentiality.

Logistics

- Provisioned through enterprise accounts.
- Data encryption and retention settings controlled by administrators.
- Often integrated with other cloud services.

Pricing

- Usage-based, usually per million tokens.
- Comparable to major vendor ZDR options but sometimes more expensive.

5 Option 5. Privacy by Proxy or VPN

Description

One idea is to route LLM traffic through a proxy or virtual private network so the model provider cannot identify the original source of the query.

Strengths

- May obscure the firm's identity from the provider.
- Simple to set up technically.

Weaknesses

- Does not prevent the provider from storing or using the text.
- Shifts subpoena risk to the proxy provider.
- Does not meet professional responsibility standards.

Logistics

- Requires contracting with a proxy or VPN provider.
- Adds latency and complexity without real confidentiality.

Pricing

- Subscription to VPN or proxy.
- Does not materially reduce LLM usage costs.

6 Option 6. Legal-Specific Services (Midpage, Harvey)

Description

Vendors such as Midpage and Harvey market LLM services built specifically for the legal industry. They promise confidentiality, legal compliance, and integrations with case law and statutes. The strategy is to use these for confidential work, and general-purpose LLMs for tasks where confidentiality is less critical.

Strengths

- Tailored to legal workflows and research.
- Strong professional-responsibility optics: designed for law.
- Often include retrieval from legal databases, improving accuracy.
- Simplifies communication with clients and courts about safeguards.

Weaknesses

- Performance may lag behind top general models for non-legal tasks.
- Two-system workflow adds friction.
- Pricing can be steep (Harvey) or more moderate (Midpage).

Logistics

- Subscription or enterprise licensing.
- Usually web-based or integrated with legal research platforms.
- Requires training staff to choose the right system for each task.

Pricing

- Midpage: likely hundreds per user per month.
- Harvey: thousands per user per month or enterprise license.
- General LLM costs still apply for non-confidential tasks.

7 Option 7. Local Anonymization with Public LLMs

Description

Another strategy is to anonymize documents locally before sending them to a public LLM, replacing names and identifiers with placeholders. The idea is that if no personal data is transmitted, confidentiality is preserved.

Strengths

- If perfectly implemented, would allow use of any public LLM.
- Useful for non-sensitive firm operations where partial anonymization is acceptable.

Weaknesses

- Effective anonymization is extremely hard.
- Combinations of non-PII data can still re-identify individuals.
- Requires a large model to do anonymization well, which undermines the goal.
- Not adequate for attorney–client privilege standards.

Logistics

- Simple version: regex and scripts; not sufficient.
- Advanced version: local LLM for context-aware anonymization; expensive and complex.
- Requires careful mapping of placeholders back to real identifiers.

Pricing

- Development time and compute resources for anonymization models.
- Still need to pay public LLM usage fees.

8 Comparative Table

Option	Confidentiality Strength	Operational Burden	Model Capability	Cost Structure	Pitfalls
ZDR with majors	High if configured	Low	Top models	Per token	Must trust policy chang
On-prem PCs	Very high	Medium to high	Limited	Upfront + electricity	Maintenanc capped mance
Hosted dedicated	High	Medium	High	Hourly GPU rental	Idle time provider log
Managed services	High	Low	Curated	Per token	Fewer choices
VPN proxy	Low	Low	Any	VPN fee + token cost	Does not storage
Legal SaaS	High	Low	Strong legal; weaker general	Subscription + tokens	Expense; w split

Local anonymization	Low to medium	High	Any (if anonymization perfect)	Token cost + local compute	Imperfect anonymizati

Looking Toward 2026

- Zero data retention will become simpler and more universal across major vendors. Expect shorter retention windows and clearer proofs of deletion.
- GPU hardware will improve, allowing larger models to fit on single chips, making on-prem and hosted options easier.
- Confidential computing features (enclaves, encrypted memory) will spread, making hosted options more trustworthy.
- Legal-specific SaaS options will expand, with more mid-priced competitors.
- Anonymization may improve through specialized legal anonymizers, but it is unlikely to become sufficient for privilege-level protection in the near term.

Conclusion

Attorneys have seven plausible paths to use LLMs without breaching confidentiality. None is perfect. The choice depends on the size of the firm, the types of matters handled, the sensitivity of client data, and the budget.

- Solo and small firms may find ZDR agreements with big vendors the most practical.
- Larger firms may combine legal-specific services with mainstream ZDR models.
- Technically inclined lawyers may experiment with hosted or on-prem hardware for local control.
- VPN proxies and simple anonymization should be avoided as inadequate.

The critical point is that confidentiality is not a single switch. It requires aligning contracts, technical settings, staff training, and professional responsibility. By understanding the seven options, attorneys can make informed choices and minimize risk.