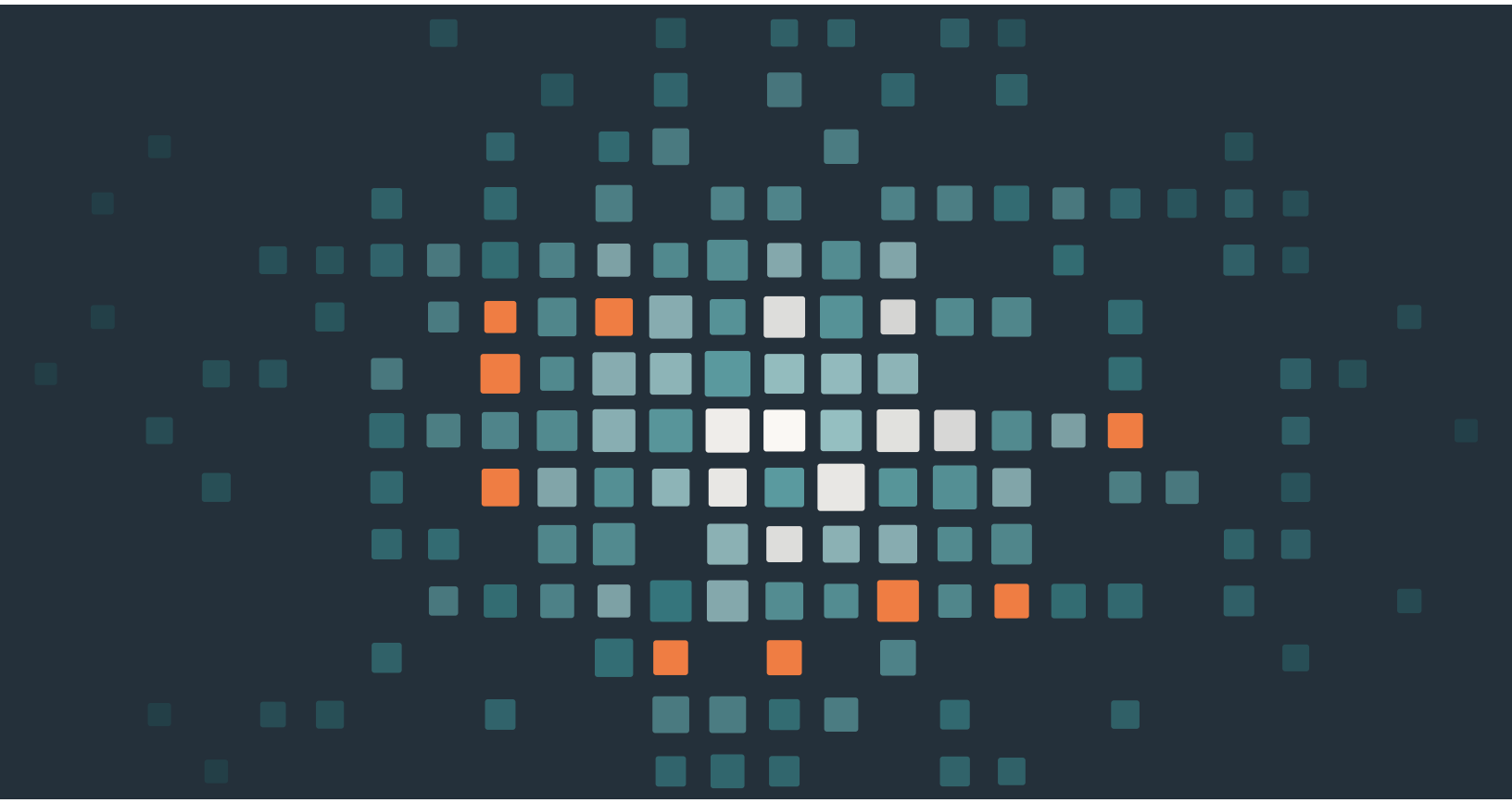
Samet Kasik

AI

Can Your Business Use AI Without Sending Its Data to the Cloud?

Local AI can keep documents off cloud services, but only if the whole workflow does. How to decide what must stay private, test a first task and compare the costs.

by Samet Kasik



KEY POINTS

- AI can run on your own computer or company server. Keeping document content there means checking the whole workflow, including document processing, search, connected services and logs.
- "Not used for training" does not mean "never leaves your business." Storage, processing, retention and access are separate questions.
- If an approved provider may process your documents, assess a managed business service before committing to local infrastructure. If external processing is prohibited, a local setup is the route to test.
- A local system earns its place when it does a worthwhile job, respects document permissions and has someone capable of maintaining it. Owning the hardware guarantees none of those things.
- Test one defined task before committing to infrastructure. Measure accuracy, source support, access restrictions, response times and time saved after human review.

Suppose you want AI to summarize a customer contract before a renewal discussion. The document contains negotiated prices, termination conditions and information you have promised to keep confidential. Uploading it would be convenient. You hesitate at the attachment button.

Could you do the same work on your own computer?

Yes. You can run a language model on a local computer or company server and process documents without sending their content to a cloud AI service. LM Studio documents offline chat and document processing; Ollama supports local models and a configuration that disables its cloud features. Models and software generally need to be obtained first, and optional online features change the data flow.[1][2]

Whether this makes sense for your business depends on the work, the data-handling rules you need to follow and your capacity to operate the system. A few confidential summaries may justify a modest desktop trial, but not a shared AI server.

What do you actually need to keep private?

“I don’t want my data going outside” can mean several things. A business may prohibit external processing altogether. Another may permit an approved provider to process information, provided it is not used for training and is deleted under agreed terms. Those requirements lead to different choices.

Separate five questions:

- **Storage:** Where are original documents, extracted text, search indexes, conversations and backups kept?
- **Processing:** Which systems receive the content to read it, search it or generate an answer?
- **Training:** Can the provider use inputs or outputs to improve its models?
- **Retention:** How long do uploads, application records and security logs remain, and how are they deleted?
- **Access:** Which employees, administrators, support providers and other parties can see the information?

For example, OpenAI says API data is not used for training unless customers opt in, while its documentation separately describes abuse-monitoring retention and storage that varies by feature. Anthropic's commercial policy also excludes training by default, with exceptions for feedback or other explicit permission. These terms apply to the specified commercial services, not necessarily to individual chat accounts.[3][4]

A service can therefore meet a no-training requirement while still processing data outside your premises. A Canadian hosting location answers a location question; provider access, jurisdiction and retention still need examination.

If an approved provider may process the documents, assess whether a managed business service meets the requirements before committing to local infrastructure. It may reduce the maintenance work the business takes on. Confirm the exact product, account terms, retention settings and connected services. Also check which accounts employees already use for work, since that determines which terms actually apply.

Local deployment also leaves questions about permitted use and safeguards intact. The Office of the Privacy Commissioner of Canada advises businesses to establish a legal basis for collecting and using personal information, limit sharing and build privacy protections into their use of generative AI.[5]

How could information leave if the model runs locally?

Follow the contract from upload to answer.

A scanned PDF may first pass through optical character recognition, or OCR, to turn images into text. A document processor extracts its contents. An embedding model creates numerical representations that help a search system locate relevant passages. Those passages are then sent to the language model with the question. This approach is commonly called retrieval-augmented generation, or RAG. It supplies reference material at the time of the question; it does not retrain the model.

Each step can run locally or use an external service. A local chat model connected to cloud OCR still sends the document outside. Locally stored files connected to a hosted embedding service may send their extracted text outside. Open WebUI's documentation explicitly separates these services and warns that its offline settings do not block all outbound connections. External APIs, web search and other configured services can remain reachable.[6]

Then examine the less visible routes: plugins, external API calls, search queries, diagnostic logs, conversation history and cloud-synced backups. An error report might contain a document extract. A web search generated from a confidential question might reveal details from it. Treat extracted text and search indexes as protected records too.

Ask whoever configures the system to show which services receive content, where copies persist and what happens when a local service fails. If the rule prohibits external processing, a failure should not silently trigger a cloud fallback. Check which model each request uses as well. An application can connect to a local Ollama server while using a cloud-hosted model through it, so a local address alone does not show where processing happens. Where external processing is prohibited, select local models and disable Ollama's cloud features.[2] Confirm the result with network controls and observed traffic, including uploads and error handling.

Installation downloads and update checks are a separate issue from transmitting customer records. LM Studio's privacy policy, for example, distinguishes local use from optional cloud services and explains the information involved in updates and model downloads.[1]

Can employees ask about documents they cannot normally open?

Consider a hypothetical shared assistant connected to an operations folder and a restricted finance folder. A coordinator can normally read service procedures but cannot open payroll records. They ask why staffing costs rose last month. If the assistant searches both folders and uses payroll details in its answer, the access controls have failed even though everything happened on an office server.

Document permissions need to govern retrieval before material reaches the model. "Do not disclose confidential information" in a prompt is an inadequate substitute. OWASP identifies unauthorized retrieval and information leakage as risks in systems that combine documents with different access restrictions.[7]

Test this with separate user accounts. Can the coordinator retrieve restricted passages, see their titles in citations or encounter them in a shared conversation? What happens when someone changes roles or loses access to a folder? Also establish who can inspect administrative logs and backups, including an outside IT provider.

For a small pilot, a deliberately restricted collection with one clear access group may be easier to manage than connecting the assistant to the entire company drive. The first question is which documents the task needs.

Will it do the work well enough?

Document search, question answering, summarization and drafting are reasonable tasks to test. In the contract example, an assistant might locate a renewal notice period, assemble relevant clauses and prepare a briefing for the person handling the account.

That person still needs to check the clause, its exceptions and any later amendment. A citation helps only when it points to the right passage and actually supports the answer. A polished summary that overlooks an amendment can create more work than it saves.

Available memory and processing power may limit which models you can run, how much document content they can handle at once and how quickly they respond. Test whether the chosen model can interpret the clauses, tables and cross-references your task depends on. If an approved cloud service is an option, run it against the same questions so the comparison reflects your own work.

Research on locally hosted RAG illustrates why the surrounding document system matters. A 2026 study found that retrieval and embedding choices affected results on a benchmark covering seven information-security standards. Its 222 multiple-choice questions did not test open-ended answer quality or citation faithfulness. The business implication is to test your own questions and documents, rather than treating a benchmark score as proof of readiness.[8]

Check whether the system acknowledges missing evidence, and compare the finished, reviewed output with the current way of doing the work, including ordinary document search.

Speed needs the same practical treatment. Measure the wait for a complete answer with usable references, then repeat with the expected number of simultaneous users. A desktop that serves one person comfortably may queue requests when several people submit long documents. Model size, document length and concurrent use affect memory and performance; Ollama's documentation describes these constraints.[2]

If checking and correcting the output consumes the time supposedly saved, the task needs a different approach.

Will owning the equipment reduce the cost?

It can, but the answer changes with utilization and the alternative you would actually buy.

A computer used for occasional summaries spreads its purchase and support costs across little work. The same machine processing a steady queue of documents has different economics. The comparison also needs comparable usefulness: a cheaper local model that cannot complete the task is not a replacement for the cloud service in the calculation.

Published savings claims deserve close reading. A preprint on consumer Blackwell GPUs separates electricity-only inference costs from hardware investment. Its break-even table assumes 30 million tokens, the units used to measure model input and output, per day. The study also identifies synthetic workloads and GPU-only power measurements as limitations. Compare your own expected volume with that assumption before relying on the result, and keep in mind that electricity per token is only one part of what the system costs to run.[9] Lenovo's 2026 comparison uses a five-year enterprise hardware lifecycle and large server configurations; its headline results should not be carried into a few-person firm's budget.[10]

Build your own comparison over the same period. Include equipment and replacement, installation, software or licences, storage, maintenance, security updates, backups, electricity and cooling, support and staff time. For cloud options, include subscriptions or usage, integration, administration, applicable storage and transfer charges, and human review. Record what happens to either estimate if usage is half the forecast. Cost per completed, reviewed briefing is a more relevant business measure than cost per generated token alone.

Someone must own the technical work from the start. During a trial, that means checking network traffic, logs, external connections and document permissions, which requires technical skill. Afterwards, it means applying updates, confirming that permissions still hold, testing model changes and restoring the service after a failure. If the business needs outside help, include those setup and support hours in the trial cost and identify who will maintain the system afterwards.

Confidentiality may justify a local system even when it costs more. The business case could be enabling valuable work that an external-processing restriction would otherwise prevent. That is a reason to pay for control, provided the work warrants it.

What would a sensible first trial look like?

Choose the trial environment after settling the data-handling requirements. Where approved external processing is permitted, test an authorized business service against the same questions and acceptance criteria. Where it is prohibited, use the local workflow below.

For a small firm, I would start with one recurring document task before buying dedicated infrastructure.

An illustrative pilot could involve two authorized employees preparing renewal briefings from a small, approved set of contracts. Run it for two weeks on existing equipment if it can handle the task. If it cannot, first establish what capacity the task needs and whether a low-cost trial is possible; buying hardware should not be the default first step. Keep the output inside the business and retain human approval of every briefing.

Agree on the test in advance:

- **The job:** Find renewal dates, notice periods and relevant amendments, with page or clause references, then draft a short internal briefing.
- **The local setup:** Keep documents, OCR, embeddings and generation on your own equipment. Disable external search, cloud features and fallback, and unnecessary integrations; confirm where logs and backups go.
- **The evidence:** Prepare a fixed set of questions with checked answers. Include scanned pages, tables, conflicting versions, missing information, superseded terms and documents one test account must not access.
- **The operating conditions:** Test normal document lengths and expected simultaneous users. Name the person responsible for setup, support and deletion of pilot copies.

Record total handling time, including corrections and review; the correctness of important facts; whether references support the answers; response times under load; and support effort. Check permission failures separately from average answer quality. A payroll disclosure cannot be offset by getting more renewal dates right.

Set acceptance criteria that fit the job. For this example, the owner might require at least a 25% reduction in preparation time after review, correct critical terms across the agreed test set, usable source references and no observed unauthorized retrieval or prohibited external transmission. These are illustrative gates for a limited trial, not a security assurance or a universal accuracy standard. The same gates apply to an approved business service; the transmission check then confirms that content reaches only the approved service and its permitted integrations.

If the system misses amendments, fix or narrow the task before expanding it. If it exposes restricted information or sends content to an unapproved service, pause its use with sensitive documents and address the cause. A successful short trial supports a next step with defined limits; it does not establish reliability across every company record.

Which setup fits your business?

Let the trial results and the level of control the work requires decide the setup.

OPTION	WHEN IT DESERVES CONSIDERATION	WHAT THE BUSINESS TAKES ON
Local computer	One or a few people need help with a defined document task, and a suitable model runs acceptably on available equipment.	Device security, local storage, updates and a workable arrangement for access and support.

OPTION	WHEN IT DESERVES CONSIDERATION	WHAT THE BUSINESS TAKES ON
Company server	A team needs shared access to controlled internal documents, with sufficient recurring use or a firm requirement for on-site processing.	User permissions, capacity, backups, availability and continuing technical ownership.
Hosted private cloud	A dedicated cloud environment offers the required controls and operating support without owning on-site equipment.	Provider and administrator access, location, contractual terms and cost. It remains external cloud processing.
Managed cloud AI	Approved external processing is acceptable and service quality, variable demand or limited technical capacity favours a managed offering.	Checking the exact product, account terms, retention settings, integrations and permitted data.
Hybrid	Different tasks have different requirements: confidential document work stays local while public research or approved work uses cloud services.	Clear routing rules and controls over what crosses the boundary, including extracts and summaries.

A hybrid arrangement can begin with separate workflows. It does not need an elaborate system that automatically chooses a model. Removing names from a document also may not remove its sensitivity; pricing, circumstances and other details can still identify a customer or reveal business information.

Whichever route you choose, preserve a way to change it. Keep original documents usable outside the AI application, document the configuration and establish whether you can replace the model or support provider without rebuilding the workflow. IBM's enterprise research treats control across data, models and infrastructure selectively; for a smaller business, that suggests protecting the dependencies that would be costly to lose or replace.[11]

Return to the contract at the attachment button. If a local assistant or an approved business service can produce a briefing you can check, handle the content according to your rules and reduce the work involved, there is a credible reason to continue. The commitment can stay small until shared use, measured demand or a stronger operating requirement makes a larger system worth maintaining.

SOURCES

1. LM Studio. *Offline Operation* https://lmstudio.ai/docs/app/offline and *LM Studio Desktop App Privacy Policy* https://lmstudio.ai/app-privacy, updated October 19, 2026.
2. Ollama. *FAQ* https://docs.ollama.com/faq, *API Introduction* https://docs.ollama.com/api/introduction and *Authentication* https://docs.ollama.com/api/authentication. Local operation, disabling cloud features, cloud models through a local server and concurrent requests.
3. OpenAI. *Data Controls in the OpenAI Platform* https://developers.openai.com/api/docs/guides/your-data?product=service. Training, abuse-monitoring logs and feature-specific retention.
4. Anthropic. *Is My Data Used for Model Training?* https://privacy.claude.com/en/articles/7996868-is-my-data-used-for-model-training, August 18, 2026. Commercial products.
5. Office of the Privacy Commissioner of Canada. *AI, Privacy, and Your Business* https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/ai_business/.
6. Open WebUI. *Environment Variable Configuration* https://docs.openwebui.com/reference/env-configuration/ and *Run Open WebUI Offline* https://docs.openwebui.com/tutorials/maintenance/offline-mode/. The offline tutorial is identified as a community contribution.
7. OWASP GenAI Security Project. *LLM02:2025 Sensitive Information Disclosure* https://genai.owasp.org/llmrisk/llm022025-sensitive-information-disclosure/ and *LLM08:2025 Vector and Embedding Weaknesses* https://genai.owasp.org/llmrisk/llm082025-vector-and-embedding-weaknesses/.
8. Jovanovski, D., et al. (2026). "An Empirical Study of Knowledge Graph-Enhanced RAG for Information Security Compliance." https://www.mdpi.com/2078-2489/17/4/389 *Information*, 17(4), 389. DOI: 10.3390/info17040389. Sections 3–5.
9. Knoop, J., and Holtmann, H. (2026). *Private LLM Inference on Consumer Blackwell GPUs: A Practical Guide for Cost-Effective Local Deployment in SMEs* https://arxiv.org/abs/2601.09527v1. Preprint, version 1. Section 4.6, Tables 9–10 and Section 5.
10. Lenovo Press. (2026). *On-Premise vs. Cloud: Generative AI Total Cost of Ownership (2026 Edition)* https://lenovopress.lenovo.com/lp2368, updated July 24, 2026. Financial framework and scenarios, PDF pp. 3–8.
11. IBM Institute for Business Value. (2026). *The Calculus of AI Sovereignty* https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-sovereignty. Printed pp. 5, 16–17 and 29. Enterprise research; its survey findings are not treated here as small-business estimates.

Web documentation reviewed November 10, 2026. Cost-study observations refer to the saved editions identified above.

Samet Kasik is Co-Founder and Business & Digital Strategy Lead at Acumen Business Consulting Inc. in London, Ontario. More essays, and a short note by email when new ones are published, at **sametkasik.ca**.

Copyright © 2026 Samet Kasik. All rights reserved.