Internal AI or Cloud AI? What Business Leaders Need to Know Before Choosing an AI Deployment Model

by David W. Koran

The financial, operational, security, and governance differences between running AI internally and paying for cloud-based AI services

August 2026

Summary

Executives frequently use the term "AI" as though it referred to one product delivered one way, but that assumption does not hold. A business can purchase AI as a cloud service, paying a provider through subscriptions, usage charges, or token-based billing. It can also operate AI models on infrastructure it controls, running open-weight models on its own workstations, servers, data centers, or private cloud environments, or it can combine both approaches.

The deployment model determines who controls the data, who operates the infrastructure, how costs scale with usage, and who accepts the technical risk when something fails. These are business questions, not engineering questions, and they belong in front of management before an AI project is approved rather than after it is deployed.

Two statements frame the entire decision.

Running AI internally does not eliminate AI costs. It converts variable provider charges into infrastructure, personnel, security, maintenance, and lifecycle-management costs.

Using cloud AI does not eliminate infrastructure responsibility. It transfers much of that responsibility to a provider and embeds the cost into subscriptions, usage charges, contracts, and vendor dependency.

Neither option is free, automatically secure, or automatically compliant. This paper explains how each model works as a business arrangement, how the economics differ, where the common misconceptions lie, and how management can select a deployment model based on the workload rather than on the marketing surrounding it. For most mature businesses, the answer will not be one or the other but a deliberate combination of both.

AI Is a Capability, Not One Product

The public experience of AI has been shaped almost entirely by consumer chat services. A user types a question, the question travels to a remote provider, the provider processes it on its own computing infrastructure, and an answer comes back. The user or the user’s employer pays a subscription or a usage charge. The provider controls the model, the infrastructure, the update schedule, and the terms of service.

That is one deployment model among several.

Modern language models exist along a spectrum of deployment options.

Proprietary cloud models are operated by providers such as OpenAI, Anthropic, Google, and Microsoft. The customer accesses the model through a web interface or an application programming interface and never touches the underlying system. The most capable general-purpose models on the market are delivered this way.

Open-weight models are models whose trained parameters are published for download. Meta's Llama family, Mistral's models, Alibaba's Qwen series, and many others fall into this category. A business can download these models and run them on hardware it controls. Open weight is not the same as open source. The weights are available, but the training data, the training process, and sometimes the permitted uses remain restricted by license. The distinction matters for legal review, but for deployment purposes the practical point is the same: the model can run on the company's own equipment.

Internal inference means operating a model on infrastructure the business controls. That infrastructure might be a single employee workstation with a capable graphics card, a dedicated AI server in a rack, a GPU appliance, space in a company data center, colocated equipment, or a private cloud arrangement. Software such as Ollama, vLLM, and LM Studio makes deployment straightforward enough that a technically competent IT department can present an internal model to employees through a familiar chat interface or an internal application programming interface.

Hybrid deployment routes different workloads to different destinations. Sensitive or repetitive work runs internally. Demanding or occasional work goes to a cloud provider.

Once these categories are visible, the executive question changes from whether the company should use AI to which workloads belong on which deployment model and what each model actually costs.

How Cloud AI Works as a Business Service

With cloud AI, the provider operates the model and the computing infrastructure. The business purchases access. That access is sold in several forms, and many companies buy more than one at the same time: per-user subscriptions for chat interfaces, metered API usage priced per token, enterprise agreements with committed spend, and bundled AI features inside existing software licenses.

The business advantages are real and explain why cloud AI dominates early adoption. Deployment is rapid and requires no infrastructure investment. The customer gets access to frontier models that no mid-market business could train or host itself.

Capacity is elastic, so a workload that doubles next month does not require a hardware purchase. The provider manages updates, security patching, and model improvements. Experimentation is inexpensive to start, and the internal technical burden is low.

The business exposures deserve equal attention. Usage-based costs are variable, and variable costs are difficult to budget. The company becomes dependent on the provider's pricing, terms, and continued operation of the specific model version it validated. Providers change models, deprecate versions, and adjust pricing on their own schedules. Data leaves the building, raising processing, retention, and jurisdiction questions that must be answered through contracts rather than physical control. Service depends on internet connectivity and provider availability. Switching providers after deep integration carries real cost. Agentic systems, discussed below, can consume tokens at rates that surprise the buyers who approved them.

The commercial structure is worth stating plainly. The cloud provider is paid based on access and consumption. The customer bears the risk of whether that consumption produces a business return. That is not an accusation; it is simply the shape of the arrangement, which management should evaluate as it does any other consumption-billed service.

How Internally Hosted AI Works

With internally hosted AI, the business operates the model itself on infrastructure it controls. A typical small deployment is a server with one or more GPUs running an open-weight model behind Ollama or a similar inference platform, presented to employees through a chat interface on the internal network. Larger deployments scale to multiple servers, dedicated appliances, or private cloud capacity.

The business advantages are concentrated in control. The company decides where data resides and who can reach the system. The system can operate without an internet connection, which matters for facilities with restricted connectivity and for continuity planning. Once the infrastructure exists, the incremental cost of one more query is electricity rather than a metered charge, which changes the economics of high-volume repetitive work. The company controls the model version, so a workflow validated against a specific model does not change because a vendor pushed an update. Integration with internal systems can be as tight as the IT department can engineer. Sensitive information can be contained within an environment the company already governs.

The business exposures are concentrated in responsibility. The company buys, amortizes, and eventually replaces the hardware. Someone on staff, or under contract,

must deploy the models, secure the system, manage access control, apply updates, monitor performance, maintain backups, and develop a recovery plan. Capacity is fixed until more hardware is purchased. Power and cooling are real line items at scale. The models that run well on affordable hardware are smaller and less capable than frontier cloud models, and the gap matters for some tasks. AI operations expertise remains scarce and expensive to hire. When the system fails at 2:00 p.m. on a Tuesday, there is no provider to call, and every failure belongs to the company.

The company may avoid a visible per-token bill, but in exchange it has assumed responsibility for operating an AI platform.

The Economics Are Fundamentally Different

The two models do not merely have different prices; they have different cost structures, and comparing them requires comparing totals rather than headline numbers.

Cloud AI is not only about tokens

The published token price is the beginning of the cloud cost, not the end. A production cloud deployment can include input and output tokens, cached-context charges, embedding generation, vector database storage and queries, agent tool calls, web search charges, image and document processing, data transfer, monitoring services, guardrail services, identity integration, developer labor, human review of outputs, and the cost of failed calls and retries.

Agentic systems deserve specific attention because they multiply consumption. A traditional chat interaction is one request and one response. An agentic workflow takes a single user request and decomposes it into a series of model calls, tool interactions, validation steps, and retries. One instruction can generate dozens or hundreds of billable model invocations. A model that looks inexpensive per token can become expensive when it runs continuously, processes large context windows, or operates as an agent. Companies deploying agentic AI should model consumption before committing, and should assume their first estimate is low.

Internal AI is not only about buying a GPU

The corresponding misconception on the other side is the executive who approves a $15,000 or $30,000 AI server and believes the company now has free AI, which is incomplete accounting.

A reasonable internal cost model looks like this:

Internal AI cost = hardware amortization + power + storage + networking + administration labor + security + maintenance + software + backup and recovery + downtime cost + human review

The cloud model should be expressed the same way:

Cloud AI cost = subscriptions + metered usage + supporting cloud services + integration labor + data services + monitoring + vendor management + human review

Notice that human review appears in both. Someone must verify AI output regardless of where the model runs, and that labor is frequently the highest unbudgeted cost in either deployment.

Where each structure wins

The structural difference is fixed cost against variable cost: internal AI is capital-intensive with low marginal cost, while cloud AI is capital-free with metered marginal cost, which means utilization decides the winner.

A workload that runs continuously at high volume, such as classifying every inbound document, summarizing every support ticket, or extracting data from every invoice, favors infrastructure the company owns, because the fixed cost is spread across an enormous number of operations. A workload that is occasional, unpredictable, or experimental favors the cloud, because paying only for what is used beats owning idle hardware. An underutilized AI server is one of the most expensive ways a company can process a small number of requests.

Data Control and Privacy

The most persistent misconception in this subject area is the belief that internal means private and cloud means exposed, and neither half of that belief survives examination.

Internal AI is not automatically private

A company can run a model on a machine it owns and still expose data through the surrounding system. Common leakage paths include cloud-based user interfaces sitting in front of the local model, external embedding services, remotely hosted vector databases, internet-connected plugins and tools, software telemetry, automatic update services, unsecured internal APIs, verbose logging, remote administration tools, backup systems that copy data offsite, and retrieval sources that themselves live in the cloud.

The fact that Ollama is installed on an internal machine does not prove that the complete AI system is private. Privacy is a property of the entire data path, not of the inference engine. A business asserting that its internal AI is private should be able to trace where prompts, uploaded files, embeddings, logs, outputs, credentials, and backups are stored and transmitted, and should verify that each of those locations is actually under its control.

Cloud AI is not automatically unsuitable for sensitive work

The reverse claim also fails. Major providers offer enterprise agreements with meaningful protections: contractual commitments not to train on customer data, configurable retention periods including zero retention for some services, regional processing options, encryption in transit and at rest, identity and access integration, and audit logging. For many categories of sensitive business information, a well-negotiated enterprise agreement with a reputable provider provides stronger practical protection than an internally hosted system administered by an overextended IT generalist.

The decision cannot be reduced to "local is safe, cloud is unsafe." The correct questions are the same in both cases: where the data goes, who can access it, what contractual or technical controls govern it, and who verifies that the controls are real. Companies subject to regulatory or contractual data restrictions must answer those questions with documentation, not with assumptions, and the answers differ by workload rather than by deployment model.

Security Responsibility Does Not Disappear

Security follows the same logic as cost: it changes shape with the deployment model, but it does not go away.

In a cloud deployment, security is shared. The provider secures the model, the infrastructure, and the service perimeter. The customer remains responsible for everything on its side: which employees can use the service, what data they are permitted to submit, how credentials and API keys are stored, how the integration is built, and whether outputs are handled appropriately. Most cloud AI security incidents in businesses are customer-side failures, including leaked API keys, employees pasting restricted information into consumer-grade tools, and integrations built without access controls.

In an internal deployment, security is direct and total. The company secures the hardware, the operating system, the inference software, the model files, the network path, the authentication layer, the logs, and the backups. An internal AI server is a

high-value system that concentrates sensitive material: prompts, retrieved documents, and generated outputs. It must be patched, monitored, access-controlled, and included in the company's existing security program. An unpatched inference server with default settings on a flat internal network provides weaker protection than a hardened enterprise cloud service, regardless of where the hardware sits.

The governing principle for management is that choosing a deployment model allocates security work rather than eliminating it.

Model Capability and Performance

Frontier cloud models are more capable than the open-weight models a mid-market business can affordably host. That gap is real, and pretending otherwise leads to failed internal deployments, but the gap is not uniform across tasks, and for a large share of routine business work it does not matter.

Internally deployable open-weight models are entirely adequate for a broad class of workloads: document classification, internal knowledge retrieval, summarization of known document types, first-draft generation, structured data extraction, tagging and routing, quality checks against defined criteria, and specialized workflows where the model is doing a narrow job repeatedly. These are also, not coincidentally, the high-volume repetitive workloads whose economics favor internal infrastructure.

Frontier cloud models retain a clear advantage for demanding general reasoning, complex multi-step analysis, broad research synthesis, sophisticated code generation, multimodal work, and open-ended tasks where the request cannot be predicted in advance. These tend to be lower-volume, higher-variance workloads, which is precisely where metered cloud pricing is tolerable.

The practical conclusion is that capability requirements should be assessed per workload. The general question of whether an internal model is good enough has no answer, but the specific question of whether it can classify the company's inbound purchase orders can be answered in an afternoon of testing.

Reliability, Continuity, and Vendor Dependence

Each deployment model fails differently, and management should know which failure it is buying.

Cloud AI fails through outages, rate limits, degraded performance during peak demand, internet connectivity loss, and provider decisions. The provider decisions are the ones that surprise companies. A provider can deprecate the model version a business validated, change pricing, revise terms of service, or alter model behavior through updates that improve average performance while breaking a specific validated workflow. A business that has embedded a cloud model deeply into its operations has acquired a dependency that should appear in its vendor risk register alongside its ERP vendor and its bank.

Internal AI fails through hardware faults, power events, storage failures, software misconfiguration, and the departure of the one employee who understood the system. There is no service-level agreement and no provider status page. Redundancy is available, but only if the company engineers and pays for it. For business-critical internal AI, that means spare capacity, tested backups, documented procedures, and more than one person who can operate the platform.

Continuity planning should treat AI like any other operational system. If the workload matters, the company needs a documented answer to "what happens when it stops," and the answer differs completely between the two deployment models.

Governance and Validation

Deployment choices attract most of the attention, but governance failures are where AI projects actually damage companies. A minimum governance program applies to both deployment models and should be in place before either one processes production data.

The program has recognizable components. An inventory of AI systems in use, including the unofficial ones employees adopted on their own. A definition of authorized uses and prohibited uses, tied to the company’s data classification so that everyone knows which information may be submitted to which system. Testing and validation of each workflow against defined acceptance criteria before deployment, with results documented. Version control, meaning the company records which model version each workflow was validated against and revalidates when the version changes, whether the change came from a provider update or an internal upgrade. Logging sufficient to reconstruct what the system was asked and what it produced. Defined human review points for consequential outputs. Periodic reevaluation, because models, providers, prices, and the company’s own workloads all change.

None of this is exotic; it is the same discipline companies already apply to financial systems, applied to a new category of system. The companies that skip it do not save time; they defer the cost to the incident that eventually forces the discipline into place.

The Hybrid Model

Most mature businesses will end up operating both deployment models, and the interesting management question is how work gets routed between them.

A workable routing logic uses four criteria. Sensitivity: information under regulatory, contractual, or competitive restriction defaults to the environment with the strongest verified controls, which may be internal infrastructure or may be an enterprise cloud agreement, depending on what the company has actually built and negotiated. Capability: tasks that genuinely require frontier reasoning go to the cloud model that provides it. Volume: continuous high-volume repetitive work migrates toward owned infrastructure as usage proves out, because that is where the economics point. Risk: business-critical workloads go wherever the company can credibly guarantee continuity.

In practice this produces recognizable patterns. Sensitive document analysis, internal knowledge retrieval, manufacturing and operational data, and high-volume extraction run internally. Complex research, occasional high-difficulty analysis, multimodal work, and experimentation run in the cloud. New use cases start in the cloud because starting there is cheap, and migrate inward if volume and sensitivity justify the infrastructure.

The hybrid model is not a compromise between two pure positions but a recognition that AI in a business is a portfolio of workloads with different requirements, and a portfolio is managed by routing each workload to the environment that fits it.

A Decision Framework for Management

The deployment decision can be structured as ten questions. No single row decides the outcome; the pattern across all ten does.

Executive question	Internal AI may be favored when	Cloud AI may be favored when
How sensitive is the information?	Data requires tight internal control	Contractual cloud controls are acceptable
How frequently will the system be used?	Usage is continuous or high volume	Usage is occasional or unpredictable
How capable must the model be?	A smaller specialized model is sufficient	Frontier-model capability is necessary
Is internet availability acceptable?	Offline or local continuity is important	Reliable connectivity is available
Is the workload stable?	Repetitive tasks can be planned and sized	Demand changes significantly

Does the company have technical staff?	Internal infrastructure can be supported	The company lacks AI operations capability
How important is model-version control?	The company requires a fixed validated model	Automatic provider improvements are desirable
How quickly must the project launch?	The company can invest in deployment	Immediate implementation is necessary
Is the system business-critical?	Internal redundancy can be engineered	Provider service levels are acceptable
How measurable is the workload?	High-volume economics justify infrastructure	The use case remains experimental

A practical selection process follows from the matrix. Define the specific workload, not "AI" in general. Classify the data the workload will touch. Answer the ten questions for that workload. Estimate total cost for each deployment model using the full cost equations in this paper, not headline prices. Test the candidate model, internal or cloud, against real examples of the work. Decide, document the decision, and set a review date, because the inputs to this decision change faster than most technology decisions a company makes.

Conclusion

There is no universal winner between internal and cloud AI, and any analysis that declares one should be treated with skepticism. The correct architecture depends on the workload.

Cloud AI shifts infrastructure responsibility to a provider and generally provides faster access to more capable models, in exchange for variable costs, vendor dependency, and data-handling questions that must be resolved contractually. Internally hosted AI gives the organization control over data, system behavior, availability, and marginal usage costs, in exchange for capital investment, operating responsibility, and a capability ceiling set by the hardware it can justify.

The businesses that get this decision right will not be the ones that picked the fashionable side; they will be the ones that understood what they were buying, what they were operating, where their data was going, and which costs and risks remained their own responsibility, and then matched each workload to the deployment model that fit it.

About the Author

David W. Koran is the founder of Efficiency Innovations, an independent publication covering how artificial intelligence is being used in business, and of David Koran & Associates, a cybersecurity and compliance consulting practice serving defense contractors and their legal counsel. He holds the CyberAB Registered Practitioner Advanced credential and ISO/IEC 27001 certifications as an auditor and implementer. He is the author of *The CMMC Decision* and *Cybersecurity in the Marine Transportation System* and writes regularly on artificial intelligence, security, and governance for business audiences. He can be reached at dkoran@davidkoran.com.