# THE BEST MODELS MUST COME TO THE DATA

## CONFIDENTIAL AI ENDS THE STANDOFF BETWEEN REGULATED DATA AND TOP-TIER MODELS


## SUMMARY

Companies with regulated data often can't send it to someone else's datacenter to use the proprietary AI models they want. Model builders, meanwhile, can't hand over the weights that make those models valuable. Both sides are protecting an asset they cannot afford to expose, and the standoff has limited what enterprises can do with their most sensitive records. Companies moved agentic workloads into production over the past two years, which raised the cost of the standoff.

Confidential AI offers another path. The model travels to the data and runs inside a protected environment. Neither the records nor the weights are exposed to anyone else, including whoever operates the hardware. That environment has to prove what it is running before either asset is released. A real enclave does four things a merely isolated workload does not. It proves itself first, protects every party from every other party, forms where the data already sits, and produces a record an auditor can check.

VAST Data built DataEnclave into a data platform its customers already run, which is a different starting point from standing up a separate confidential stack. Records stay in place, weights decrypt only inside an attested enclave, and every attestation and key release lands in a queryable record. For regulated buyers, the technology question is largely settled. The open questions are which shelved projects are worth reopening and what proof to demand from any vendor selling an enclave.


## THE STANDOFF THAT STALLS ENTERPRISE AI

Enterprise AI used to work in a simple loop. Someone asked a question, a model answered, and a person decided what to do with the answer. Agentic AI takes that person out of the loop. Agents run on their own, reach across systems, and hand work to other agents without pausing for approval. These agents already summarize wikis, triage ticket queues, and write code against internal repositories.

But running a top-tier AI model on regulated data brings together three parties, each with something it cannot hand over:



Copyright © 2026 Moor Insights & Strategy

- **The regulated enterprise** holds data that it's obligated to keep inside an environment it controls. The company can share that data with outside providers it has reviewed and approved, but sharing it elsewhere can violate regulations and expose the company to fines.
- **The model builder** holds weights that took years of research and an enormous amount of computing power to produce. Anyone who gets a copy can run the model themselves, and there is no way to take that copy back. So the builder hangs onto them for competitive reasons.
- **The infrastructure operator** runs the hardware where the model and the regulated data meet, so it can see what executes on that hardware. That access comes with the job, and it simply cannot be avoided.

Many companies play two of these roles. Consider a bank that fine-tunes its own model on its own customer records. It holds regulated data and valuable model weights at the same time — and needs both of them protected from whoever runs the hardware.

What blocks the bank is not model quality. Today's open-weight models handle many agentic tasks well enough. What stays constrained is access to certain classes of proprietary and regulated records. Underwriting files, insurance claims, patient records, and financial trading positions are subject to rules that don't bend. Running those records through a proprietary model the company doesn't control means someone else decides who can see them.

An agent could read a claims file as easily as a wiki page. The technical work is identical. But those records stay off limits because of what happens when the contents get out. Exposing an internal wiki page costs a company some embarrassment and little else. Exposing a claims file is a different order of problem. The information inside belongs to real people, and once it has been exposed, there is no getting it back.

Enterprise AI programs that need access to sensitive data eventually run into the same wall, and the wall is legal and reputational rather than technical. Nobody wants to lose their customers' or stakeholders' trust by exposing their data, or their model, to a party they do not control.

## *Encryption Stops Working the Moment Data Gets Used*

For decades, companies have encrypted data on disk, then encrypted it again as it crosses the network. Security calls those two states "data at rest" and "data in flight." A third state, "data in use," has never had the same protection. Outside of specialized


Copyright © 2026 Moor Insights & Strategy

research techniques, a processor has to unscramble data before it can work on it, whether that processor is a CPU or a GPU. That readable copy then sits in memory for as long as the job takes.

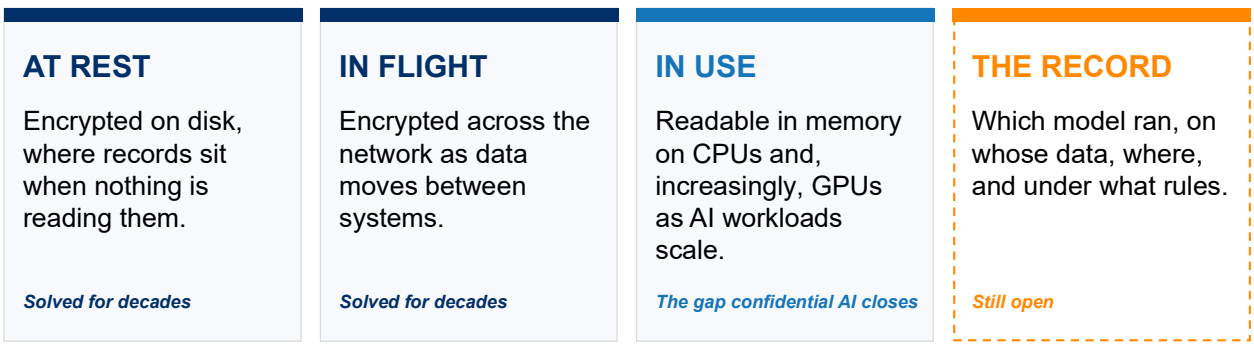
At the moment the data is most useful, therefore, it sits in readable form — often on hardware that someone else operates. Whoever operates that hardware might have no permission to see the data. This is why regulated industries run a great deal of AI, but only on models they can host themselves. Meanwhile, most of the strongest models are available only as an outside service.

Confidential AI is the emerging approach to protecting data from that exposure. Confidential computing, the hardware technique underneath confidential AI, arrived nearly a decade ago to keep data unreadable to whoever runs the machine. A processor walls off a region of memory that nothing else on the system can read: not the operating system, not the software managing the virtual machines, and not the company running the hardware.

CPUs delivered that capability first. As companies moved AI into production, those GPU-based workloads needed the same protection, so GPUs have run using a confidential mode since 2022. What changed more recently is that turning on the protection barely slows the work down. That protection now reaches the connections between chips, not only the chips themselves, which is important because a big model runs across many GPUs at once, and the data moving between them used to travel unprotected. The protection now reaches everything an AI workload touches, from where the data sits to what gets recorded once the job is done. What's new is that data owners, model providers, and infrastructure operators all work inside one runtime.

Figure 1 shows where protection stands today.

FIGURE 1: HOW FAR PROTECTION REACHES AS DATA MOVES THROUGH AN AI JOB




*(Source: Moor Insights & Strategy)*


Copyright © 2026 Moor Insights & Strategy

Sovereignty and sovereign AI are discussed alongside confidential AI often enough that the terms can blur together. Sovereign AI is about where data physically sits and who holds legal authority over it. Confidential AI is about whether anyone can read that data while a model works on it. Most sovereignty programs need both.

## WHY MOVING THE DATA STOPPED WORKING

Almost every company runs AI the same way today. The data travels to wherever the model lives, usually a service someone else operates. The arrangement works well enough for most workloads. It falls apart the moment the data is regulated, because sending the data out is precisely what the regulations prohibit.

The result is usually an AI program split cleanly in two. On one side sit the projects using unrestricted data, which can use whichever model is best. On the other side sits everything that touches regulated records, and those projects are limited to models the company can run itself.

The regulated half of that program is getting harder to ignore. In many cases, teams have already built what they can on unrestricted data. The restricted data, the records too sensitive to release, tends to be the most valuable. A healthcare company gets more out of a model that reads granular patient data. A bank gets more out of one built on specific account details. That information sits at the center of the business, which is the same reason the rules around it are strict. So the projects stuck behind that data often promise the biggest returns on a company's AI roadmap.

A project that puts regulated data into a top-tier model goes to security review before it can run. Reviewers check it against the company's own security requirements, and most projects fail there. Hosting the model inside the company would solve the problem with the data, but that requires GPU capacity and the people to integrate it, which is a project in itself. Even then, a company still needs a copy of the weights from the model builder, and that rarely happens. Some teams also rule out open-weight models on provenance grounds, which narrows the in-house option further.

### *THREE WORKAROUNDS AND THEIR TRADEOFFS*

Teams that can't use their preferred AI models with regulated data typically turn to three workarounds. Each addresses the access problem but introduces a tradeoff that shifts the cost elsewhere rather than removing it.


Copyright © 2026 Moor Insights & Strategy

### TABLE 1: WORKAROUNDS AND TRADEOFFS

| THE WORKAROUND | THE TRADEOFF |
|---|---|
| **Masking and redaction.** Sensitive fields get stripped or substituted before the request goes out to an external model. | The model answers a question about a record stripped of its most useful details. The answer comes back weaker, sometimes too weak to be worth the effort. |
| **Smaller models hosted locally.** A smaller model with published weights runs inside the company's own datacenter, so it can see real records. | The smaller model isn't as capable as the model the team wanted, and the bigger model's capability was the reason the project got funded. |
| **Deferral.** The project goes into a review queue instead of getting cancelled outright. | The executive sponsor moves on, the queue never clears, and eventually nobody can say what happened to it. Deferral is the most common outcome, and the one that nobody logs. |

## THE ENCLAVE AND WHAT IT HAS TO PROVE

Bringing the model to the data is the obvious answer, because it allows the model to run on whatever platform already holds the records. The operator might be a cloud provider, a partner, or the company's own infrastructure team. The hardware might sit in a cloud or in the company's own datacenter. Moving the model there allows the operator to hold both the regulated data and the model weights at the same time.

An enclave is a place where the data owner, the model builder, and the infrastructure operator can work together without any of them seeing what the others brought. The enterprise brings the data it isn't allowed to release. The model builder brings the weights it isn't willing to release; the model reads the data to perform inference, but its builder never sees any of it. The operator runs the job without being able to inspect either the data or the model.

### *FOUR TESTS AN ENCLAVE HAS TO PASS*

Four requirements separate a real enclave from a workload that is merely isolated. They also provide a framework for evaluating any vendor selling one.

1. **Nothing is released until the environment proves itself.** The hardware produces a signed description of what's running, which is called attestation. The signature traces back to the chip vendor, so the company operating the hardware can't forge or alter it. The records and the weights entering that environment are protected assets belonging to different owners. Each asset owner can independently verify the environment before releasing the keys or secrets required to access its protected asset.


Copyright © 2026 Moor Insights & Strategy

2. **Everyone is protected from everyone else.** Weights stay unreadable to the company that owns the data. Records stay unreadable to the model's builder, even while the model is working on them. The operator is unable to inspect either one.
3. **The enclave forms where the data already sits.** An enclave has to work wherever the company operates, including sites with no outside connection at all. Sovereignty and defense programs require exactly that, but many different use cases also benefit from reducing the distance traveled, and therefore the response time, for the many thousands of small reads common to agentic enterprise workloads.
4. **It produces a record.** Nothing inside an enclave can be inspected, which is the entire point, so the usual security tooling can't see inside it either. Compliance teams still have to account for which model ran, whose data it touched, where it ran, and under what rules. That information has to be available on demand. Stitching log files together after the fact doesn't work.

## How VAST Data Approaches Confidential AI

VAST Data has spent roughly a decade building a data platform specifically for AI work. Storage, a database, and a compute runtime all sit inside one system. According to VAST, a quarter of the Fortune 100 uses the platform, including companies in regulated industries, and so do a majority of the neoclouds. Neoclouds are the infrastructure operators that a regulated company would usually rent GPU capacity from, so VAST sits underneath both the enterprise and the cloud it rents.

As part of VAST's platform, a secure container runtime already ships to customers. VAST introduced VAST DataEnclave, the confidential AI capability of the VAST DataEngine, to turn that runtime into a confidential execution environment. Jobs run in confidential virtual machines across supported CPU architectures, with GPUs in confidential compute mode; attestation verifies the environment before anything protected is released into it. For VAST customers, DataEnclave is a capability that can be switched on inside a platform they already run, which costs less to adopt than standing up a second stack.

### *What a Regulated Buyer Gets with VAST DataEnclave*

The architecture splits into three parts that VAST calls planes. Records live in the secure data plane and stay there until a job needs them. Models load and run in the confidential workload plane. Attestation and key handling sit apart in the external trust


Copyright © 2026 Moor Insights & Strategy

plane. When a job starts, the records load into the enclave. The model loads into the same enclave, where the GPU works on both. DataEnclave provides the attestation and verification framework, while data owners and model providers retain control over the policies and key infrastructure governing access to their respective assets. Nothing is released until the environment proves itself. The signed description of what's running has to match the values an owner approved in advance, and a mismatch means no keys are released.

Protection runs in every direction. Weights decrypt only inside the attested enclave, so the company running the job never holds them. By VAST's account, data never leaves the customer's environment and is readable only inside the enclave, so the model processes it without the model builder ever gaining access. Keys to the weights stay with the model builder.

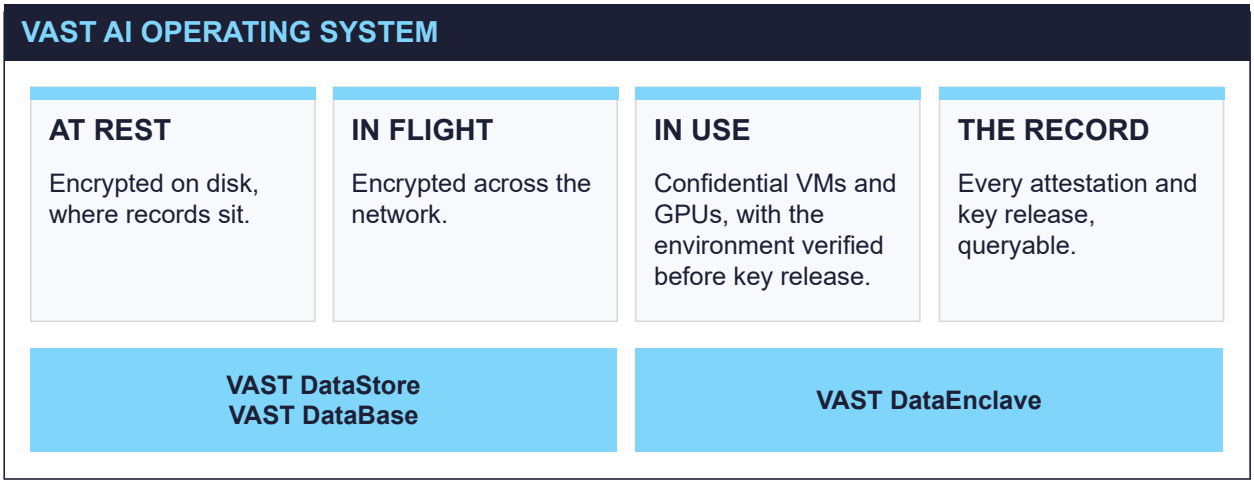
The enclave forms around the data, wherever that data lives. Nothing gets copied to another system. The enclave works at edge sites, in private clouds, in AI clouds, and on appliances from server vendors. The DataEnclave attestation service is built on the open CNCF Trustee stack for connected deployments, and fully air-gapped sites get attestation and key brokering on-premises instead. Models that a company trains itself deploy the same way. Tuned weights travel wherever they're needed, without exposing the tuning that made them valuable.

Enclaves protect data while the model works on it. They do not, by themselves, produce the record of what ran. DataEnclave writes that record into the VAST DataBase, capturing every attestation decision, key release, policy decision, and enclave lifecycle action. VAST describes that record as immutable and cryptographically verifiable, so an auditor can independently verify it. Teams can query that history while a job is running, or long after it finished. Taken together, the runtime, the attestation service, and the immutable record allow enterprises, model builders, and infrastructure providers to each build what VAST describes as a confidential AI factory.

Figure 2 maps the same four states onto the VAST AI Operating System. The platform already covered data at rest and in flight before DataEnclave existed. DataEnclave adds protection for data in use, and the record it writes into the VAST DataBase closes the fourth. All four states now sit inside one platform.


Copyright © 2026 Moor Insights & Strategy

FIGURE 2: HOW THE VAST AI OPERATING SYSTEM COVERS ALL FOUR STATES




*(Source: Moor Insights & Strategy)*

## THE DECISION IN FRONT OF AI BUYERS

The data a company most wants to put in front of an advanced AI model is usually the data it can least afford to send out. Some of that data is regulated and some of it is simply proprietary, but either way, it needs to stay home. Yet the strongest models are delivered as an outside service. Bringing one in-house is not the customer's call, because the model builder decides who gets a copy of the weights.

Confidential AI puts the job inside an enclave instead, where each party keeps control of its own assets. VAST's version is DataEnclave, an enclave built inside a data platform customers already run. GPU capacity is still required, whether the company owns it or a partner operates it, but the integration work of making that capacity confidential is already done.

### *FOUR QUESTIONS AI BUYERS SHOULD ASK*

A company wanting to use powerful AI models on data it cannot send out should start with four questions:

1. Which AI projects got shelved because the data was too sensitive to hand over, and not because the concept of the project was weak?
2. Which of those shelved projects might change the business if a top-tier model ran against the real, full records?


Copyright © 2026 Moor Insights & Strategy

3. What is the company already giving up to work around data it can't send out? Count the weaker models, the longer timelines, and the projects that never shipped.
4. Can the vendor show what record an auditor would receive a year from now and explain who produces it? The answer separates a real confidential AI architecture from an isolated workload with a new name on it.

Deciding which systems an agent can reach, and which actions it can take without a human, remains a job for the business. What changes with confidential AI is that the data and the weights are protected from the environment running them. That environment has to prove itself, cryptographically, before either one is released. Companies no longer need to choose between the data they're obligated to protect and the models they actually want. The model travels to the data now, and it arrives in a location the company chooses. With that protection in place, it's time to go back to the AI projects that stalled.


Copyright © 2026 Moor Insights & Strategy

# Important Information About This Paper

## Contributor

Mike Leone, Vice President and Principal Analyst, Enterprise Data Platforms, Analytics, and Governance

## Publisher

Patrick Moorhead, CEO, Founder, and Chief Analyst at Moor Insights & Strategy

## Inquiries

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

## Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying the author's name, author's title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

## Licensing

This document, including any supporting materials, is owned by Moor Insights & Strategy. This publication may not be reproduced, distributed, or shared in any form without Moor Insights & Strategy's prior written permission.

## Disclosures

VAST Data commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

## Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.

© 2026 Moor Insights & Strategy. Company and product names are used for informational purposes only and may be trademarks of their respective owners.