ARTHEUS

WHITE PAPER

DEFENSE AI · DATA SOVEREIGNTY

AI Behind the Boundary

Deploying modern AI where the data cannot leave — a practical architecture for on-premises transcription, translation, and analysis in classified, controlled, and disconnected environments.

 Intelligence. Integrity. Impact.

artheus.co

Defense AI · Data Sovereignty

Ask a program office whether it wants what a modern language model can do — summarize a night's worth of field reports before the morning brief, translate source material across forty languages, pull the three paragraphs that matter out of a four-hundred-page document — and the answer is almost always yes. Ask whether the underlying data can be shipped to a commercial cloud service to get it, and the answer is almost always no. That gap is where most federal AI ambitions quietly stall. This paper is about closing it without moving the data — and about the engineering that makes the resulting system one you can actually trust, audit, and defend.

The problem, in three numbers

$4.88M

Average total cost of a data breach in 2024 — the price of data ending up where it shouldn't.

IBM, Cost of a Data Breach 2024

~80%

Share of enterprise data that is unstructured — audio, documents, images, and free text, the material AI is best at and cloud rules are strictest about.

IDC, widely cited industry estimate

0

Bytes that leave the boundary in a correctly built on-premises architecture — the number this paper is written around.

Artheus design principle

Data sovereignty is a boundary, not a preference

A large share of the government's most valuable data is, by rule, not allowed to leave the environment it lives in. Controlled Unclassified Information carries handling obligations under DFARS 252.204-7012 and NIST SP 800-171. Classified systems answer to ICD 503 and never touch the commercial internet. Personal and operational data brings its own constraints, and forward or disconnected operations remove the network entirely. For any of these, the default commercial pattern — send the prompt and the data to a vendor's API, get an answer back — is simply off the table.

This is not risk-aversion, and treating it as a hurdle to be waived misreads the problem. The boundary is the requirement. The useful question is not "how do we get an exception," it is "how do we get the capability without asking for one." And the stakes are not abstract: when sensitive data does leak, the cost is measured in millions and, for national-security data, in terms that do not fit on a balance sheet (Figure 1).

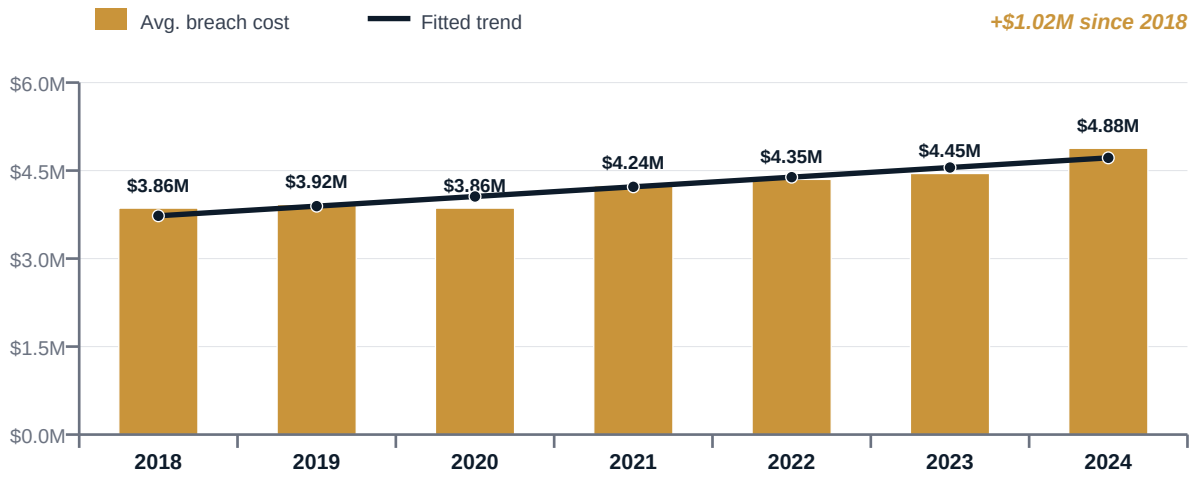

Figure 1. Seven years of rising breach costs, with the fitted trend underscoring the direction — the global average reached $4.88M in 2024, more than $1M above 2018. When data cannot leave the boundary, the cheapest breach is the one that never had the chance to happen. (Source: IBM, Cost of a Data Breach, 2018–2024.)

Why the usual answers fall short

- **“Wait for an accredited cloud AI service.”** Authorization timelines run in years, and even a fully accredited multi-tenant service still asks the most sensitive data to leave its enclave and sit next to someone else’s. For the data that matters most, that is the wrong trade on the wrong schedule.
- **“Don’t use AI on sensitive data.”** This leaves the highest-value material — the sensitive material — the least served. The workforce ends up with AI for the unclassified email and nothing for the mission.
- **“De-identify, then send it out.”** Masking is lossy and imperfect, re-identification is a live risk, and for classified data it is a non-starter. A redaction step does not convert a sovereignty problem into a convenience.

> *The instinct to wait for someone else’s accredited cloud is understandable, and for data that by definition is not allowed to leave, it is usually the wrong bet.*

A different architecture: run the whole pipeline in-boundary

The premise that made cloud AI feel mandatory — that only a hyperscaler can run models this capable — is no longer true. Speech recognition, machine translation, optical character recognition, and mid-size open-weight language models now run on a single well-specified GPU server or workstation, air-gapped if the mission calls for it. The architecture keeps every stage inside the boundary, from raw input to grounded, audited output (Figure 2).

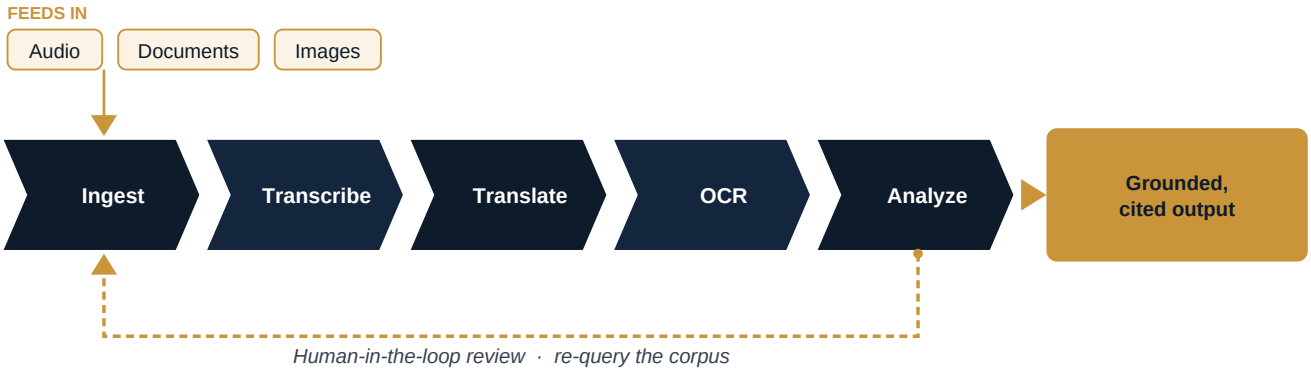

Figure 2. The on-premises pipeline as a closed loop. Multilingual audio, documents, and images feed in; local models transcribe, translate, read, and analyze; the output is grounded in its sources and written to a tamper-evident audit chain. A reviewer verifies and re-queries — and at no step does anything leave the box.

None of this requires a data center. It requires one properly specified machine, an engineering discipline for keeping models and data in-boundary, and an honest accounting of what the hardware can and cannot do. The constraint is real — a local deployment will not match the raw scale of a frontier cloud model — but for the tasks that matter here (transcribe, translate, read, retrieve, summarize with citations), the gap is narrow and closing, and it is a price worth paying for data that cannot leave.

Cloud versus boundary: the trade for sensitive data

For unclassified, low-sensitivity work, commercial cloud AI is often the right call. The calculus inverts the moment the data cannot leave. Figure 3 lays the two options side by side on the dimensions that decide a sensitive deployment.

Dimension	Commercial Cloud AI	On-Premises AI (in-boundary)
Data leaves the boundary	Yes — sent to a vendor	No — never leaves
Usable for classified / air-gapped	No	Yes
Time to field on sensitive data	Months to years (accreditation)	Weeks
Provenance & tamper-evident audit	Vendor-dependent, opaque	Built in, inspectable
Works disconnected / at the edge	No	Yes
Raw model scale	Highest	High and rising

Figure 3. For data that cannot leave, the trade-offs favor the boundary on every dimension except raw scale — and for transcription, translation, and grounded analysis, scale is rarely the binding constraint.

Trust is a property you engineer, not a claim you make

Running a model locally solves where the data goes. It does not, by itself, make the output usable for a decision or an audit. Three properties do, and they are the difference between a chatbot and an evidence system:

- **Citation-grounding.** Every finding points back to the specific source span it came from, so a reviewer checks the receipt rather than trusting the summary. This is what lets a language model support evidence, not just draft prose — and it is the single most important defense against a confident, wrong answer.
- **A tamper-evident audit trail.** A hash-chained record (SHA-256) of inputs, model version, and outputs means that if evidence or findings are altered after the fact, the chain breaks and it shows. Oversight gets provenance, not assurances.
- **PII masking and reproducibility.** Personal data is protected while analytic value is preserved, and pinned model versions make a result inspectable and repeatable — an advantage a constantly-updating cloud endpoint cannot offer.

These are not features bolted onto the end. They are the reason a leader, an auditor, or an inspector general can rely on what the system says — and the reason the same architecture that satisfies the security office also satisfies the mission owner.

It lines up with where governance is going

The NIST AI Risk Management Framework organizes AI governance around governing, mapping, measuring, and managing risk. An in-boundary, citation-grounded, audited architecture maps onto that cleanly: the data never leaves, provenance is measurable, and a human stays in the loop by design. Federal policy is pushing in the same direction from the adoption side — OMB M-25-21 (2025), issued under Executive Order 14179, directs agencies to accelerate AI use while standing up real governance and public trust, including risk management for high-impact systems. On-premises, auditable AI is how an agency says yes to adoption without failing the governance test. The Department of Defense’s data and AI adoption strategy makes the same point from the mission side: decision advantage at the edge, where connectivity is poor and the data cannot move, is exactly where the need is sharpest.

A pragmatic path to standing it up

The failure mode is not technical, it is scope. Teams try to boil the ocean, stall in accreditation, and conclude on-premises AI is too hard. It is not, if you start narrow and prove it:

- **Pick one bounded, painful use case.** A backlog of foreign-language recordings; a document review that eats analyst-weeks. One workflow, one clear before-and-after.
- **Right-size the hardware to that case.** A single GPU workstation or server, specified to the workload, not a speculative cluster.
- **Build the trust properties in from day one.** Grounding and the audit chain are not a later phase; they are what makes the pilot’s output admissible.
- **Measure against the manual baseline,** then expand to the next workflow on the same machine and the same controls.

Where it pays off

- **Third-party monitoring and oversight** — independently verifying that programs, contracts, and partner activities happened as reported, with an evidence trail that holds up under scrutiny.
- **Multilingual field-report and open-source triage at the edge** — turning foreign-language reporting into usable analysis where time is short and the network is not there.
- **Sensitive-environment analysis** — bringing modern AI to the data that genuinely cannot leave, instead of leaving that capability unused.
- **Accountability and audit** — giving inspectors, auditors, and oversight functions a chain of evidence they can verify independently.

The boundary is not going away

Data-handling rules are tightening, not loosening, and the volume of multilingual, unstructured mission data is climbing. The organizations that come out ahead over the next few years are the ones that stop reading “we can’t send the data out” as a reason to go without AI, and start reading it as an architecture requirement they can actually meet. The technology to run capable AI entirely inside the boundary exists today. What remains scarce, and what decides whether the output can be trusted, is the discipline to build it so that every finding is grounded, every step is audited, and every result can be defended.

References

1. National Institute of Standards and Technology, *Zero Trust Architecture*, NIST Special Publication 800-207, August 2020.
2. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, January 2023.
3. Office of Management and Budget, *Accelerating Federal Use of AI through Innovation, Governance, and Public Trust*, OMB M-25-21, April 3, 2025 (issued under Executive Order 14179).
4. U.S. Department of Defense, *Data, Analytics, and Artificial Intelligence Adoption Strategy*, 2023.
5. IBM Security, *Cost of a Data Breach Report*, 2022, 2023, and 2024 editions.
6. Defense Federal Acquisition Regulation Supplement, clause 252.204-7012, *Safeguarding Covered Defense Information and Cyber Incident Reporting*; NIST Special Publication 800-171.
7. Office of the Director of National Intelligence, *Intelligence Community Information Technology Systems Security Risk Management*, ICD 503.
8. U.S. Department of Defense, *Cloud Computing Security Requirements Guide (SRG)* — Impact Levels 2 through 6.