Approaches to governing AI, compared

light

$670K

Shadow AI, the unsanctioned AI tools an institution did not approve, added about $670,000 to the average data breach. The same report puts the global average breach at $4.44 million. Source: IBM Cost of a Data Breach Report 2025.

> The hardest case is not the AI you approved. It is the AI you did not: the personal accounts, browser sessions, and command-line clients that never touch a sanctioned path. A control that governs only what was wired to it cannot see this. The job is to decide where AI traffic leaves the device, before it leaves, across the AI tools that reach a provider, no matter which tool sent it.

This is not a hypothetical. The PowerSchool breach disclosed in early 2025 turned on a single set of stolen credentials reaching a system that held student records; the threat actor claimed data on roughly 62 million students, a figure PowerSchool did not confirm (source: BleepingComputer, January 2025). The lesson for AI is the same. The exposure lives on the path no one was watching, and the institution owns the consequence whether or not it sanctioned the route.

The four questions position decides

WHAT POSITION HAS TO ANSWER

AND THE QUESTIONS UNDER THEM

- **Decides at the device.** Does it decide before anything leaves, across the AI tools that reach a provider, rather than only the ones configured to cooperate?
- **Evidence you can verify.** Does it produce a record you can show was not altered, verifiable on its own terms rather than on the vendor's word?
- **Custody.** Do the keys and the content stay with you, so no outside operator becomes the holder of what your staff sent?
- **Data provenance.** Does the record span the providers and accounts your staff use, accounting for what was sent, what was redacted, and where it went?

The common approaches, and where each leaves a gap

Acceptable-use policy alone	A written rule about what staff may and may not do with AI. The gap: a document does not enforce anything. It cannot see use, redact a prompt, or produce a record, and unsanctioned use continues unobserved.
Network or DNS blocking	A firewall or DNS filter that allows or denies whole destinations. The gap: the decision is all or nothing. Block a provider and staff route around it through another tool or a personal device. Allow it and everything passes unseen. It cannot read inside an encrypted session, redact, or keep a usable record.

An API gateway that requires integration	A proxy that applications are configured to send their AI calls through. The gap: it governs only the traffic that was wired to it. A new tool, a personal account, a browser, or a command-line client that was never configured passes around it. It also tends to terminate the session centrally and see plaintext, which makes the operator a holder of your data. It does not govern what it was never told about.
A closed or walled-garden assistant	An enterprise AI suite that governs its own sanctioned assistant. The gap: it governs one tool. The moment a user opens a different provider, it is blind. Coverage equals a single application, not the institution.
A browser extension or per-app plugin	A control installed into one surface, usually the browser. The gap: it covers that one surface. Desktop apps, IDEs, command-line clients, and other browsers pass by, and a user can often turn it off.
Data-centric DLP or CASB	Tools built to watch files and SaaS usage, often by sampling. The gap: they are not aware of the AI decision itself. They observe data movement, often after it has already happened, rather than deciding before content leaves the device.
Provider-side controls	Relying on the model provider's own settings and logging. The gap: by the time those controls apply, the data has already left your boundary. The record belongs to the provider, not to you. It is not independent or tamper-evident, and it does not span the other providers your staff use.

Where Verillian sits

Verillian sits at the device, before content reaches a provider. Because the control sits where traffic leaves the machine, it can answer all four questions rather than assume the rest away.

Decides at the device	It sees the AI traffic leaving the device across the AI tools that reach a provider, regardless of the application, the provider, or whose account it is. Nothing has to be wired up to be seen, and policy is applied at the moment of execution: allow, redact, or block. When a decision is uncertain it fails closed, and a blocked request does not leave the device. Detection and redaction of regulated data are best-effort, not a guarantee that every sensitive value is caught.
Evidence you can verify	Every captured interaction is Ed25519 signed and SHA-256 hash-chained into a record your institution holds. Any change to an entry breaks the chain, so the record is tamper-evident and verifiable mathematically, independent of vendor attestation, rather than taken on trust.
Custody	Content is encrypted under your institution's key. The server stores ciphertext it cannot read, so no outside operator becomes the holder of what your staff sent, and even Verillian cannot read your content.
Data provenance	One control spans the providers your staff use, including air-gapped operation, so the record accounts for what was sent and what was redacted across the tools they actually reach. Audit content is fully parsed for Anthropic, Claude and Claude Code today, with best-effort detection across other providers, expanding to the providers your institution uses.

THE SAME QUESTIONS, ANSWERED SIDE BY SIDE

APPROACH	SEES UNSANCTIONED AI	NO INTEGRATION NEEDED	DECIDES BEFORE EGRESS	REDACTS, NOT JUST BLOCKS	TAMPER-EVIDENT EVIDENCE	SPANS PROVIDERS USED
Acceptable-use policy	no	yes	no	no	no	no

APPROACH	SEES UNSANCTIONED AI	NO INTEGRATION NEEDED	DECIDES BEFORE EGRESS	REDACTS, NOT JUST BLOCKS	TAMPER-EVIDENT EVIDENCE	SPANS PROVIDERS USED
Network / DNS blocking	partial	yes	partial	no	no	partial
API gateway (integrated)	no	no	yes	partial	partial	partial
Closed assistant	no	yes	partial	partial	no	no
Browser / app plugin	partial	yes	partial	partial	no	no
DLP / CASB	partial	partial	no	partial	partial	partial
Provider-side controls	no	yes	no	partial	no	no
Verillian	yes	yes	yes	best-effort	yes	yes

Read the table by column, not by row. **Yes** means the approach meets the question, **partial** means it does so only conditionally or for part of the surface, **best-effort** means redaction runs before egress but does not guarantee every sensitive value is caught, and **no** means it does not. The pattern that emerges is not that one tool is better at everything. It is that position decides which questions a control can answer at all.

The point

Where a control sits decides what it can catch. Sit downstream of the decision and you inherit every blind spot upstream of it. Most approaches govern the AI you already approved. The exposure lives in the AI you did not, and breach costs are highest where regulated data sits, $7.42 million on average in healthcare, the costliest industry, and $5.56 million in financial services (IBM Cost of a Data Breach Report 2025). A control at the device, deciding before egress, with a record you hold and can verify, is the position that answers all four questions instead of one.

Sources: IBM Cost of a Data Breach Report 2025 (global average $4.44 million; shadow AI added about $670,000; healthcare $7.42 million; financial services $5.56 million). PowerSchool student-data breach, threat-actor claim of roughly 62 million students, reported by BleepingComputer, January 2025. The software is source-available under the Elastic License 2.0, so your security team can read exactly what runs on each machine.

The model proposes. Verillian decides, and gives you the receipt.

If you are evaluating controls, we will walk this table against your environment and show the signed record on your own infrastructure, under your own keys.

hello@verillian.ai