WR
White Rock Future Tech
Little Compton, Rhode Island | Leesburg, Florida

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

SOLUTION PAPER

VPNs and TLS Protect the Pipe, Not What AI Providers Keep

An owner, director or board member who has already decided AI is useful and now has to decide what their staff may send where. It is the companion to What You May Not Know About Intellectual Property And AI, which answers why this matters. This paper answers how.

October 2026. Provider terms and product names change faster than any paper can track, so every claim here carries a numbered source you can follow and re-check. Where we could not confirm something from a primary source, the paper says so rather than rounding it up into a fact.

SECTION 1

What this paper is for

The question we get asked is some version of: our people are already using these tools, how do we stop the company's own information walking out the door?

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

The honest answer has three parts, and only the third one is about technology.

1. Decide what may never leave the building. Most organizations have never written this down.
2. Pick a destination for each kind of information, and accept that the contract you signed decides more than any setting you can toggle.
3. Build the controls that match. Some of the popular ones do nothing for this problem.

This paper walks all three. It names the controls that work, the controls that help with something else, and the controls that are sold against this threat and do nothing about it. We would rather tell you which of these not to buy.

SECTION 2

The sentence everything else follows from

Network controls decide who may talk to which service, from where, and whether anyone can watch along the way. They decide nothing at all about what the service does with the data once it arrives.

Almost every expensive mistake in this field is a control from the first half of that sentence, sold against a problem in the second half.

The clearest example is encryption in transit. TLS 1.3 does exactly what it claims: it defeats an eavesdropper and a tamperer on the path between you and the service. [1] But the AI provider is not on the path. The provider is the destination. A flawless

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

connection is a reliable way of delivering your client's data into somebody else's plaintext. "We use TLS 1.3" is not an answer to "what happens to our data".

The documented default at the other end is a stored copy. OpenAI retains abuse-monitoring logs for 30 days, and says those logs may contain prompts, responses and derived metadata.[2] Anthropic deletes within 30 days on the API but keeps flagged conversations up to two years and safety classifier scores up to seven.[3] These are not scandals. They are published terms. The point is that they are decided by the contract, not by your firewall.

SECTION 3

Step one: write down what may never leave

This is the step people skip, and skipping it makes every later decision arbitrary.

You need categories, and each category needs a destination. Four tiers cover most commercial organizations:

- **Public.** Already published or intended for publication. Any destination.
- **Internal.** Ordinary business information. Embarrassing if leaked, not damaging.
- **Confidential.** Customer records, personal data, financials, contracts, pricing.
- **Restricted.** The things that are the business: designs, source code, formulations, proprietary processes, legal strategy, anything under a duty of confidence to a third party.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

Public bodies have the categories handed to them already, through the Controlled Unclassified Information program and the FIPS 199 impact levels, and should use those rather than inventing a parallel scheme.

The rule that makes the scheme real: **the category decides the destination, and a person who does not know the category cannot pick the destination.** That is why the policy has to be written in terms of the work (“quotes”, “patient notes”, “the CAD files”) rather than in terms of abstractions. A classification scheme nobody can apply at the keyboard is decoration.

SECTION 4

Step two: the ladder of destinations

Ranked by how much it genuinely protects, strongest first. The reasoning for each is in the sections that follow.

Destination	What it protects against	What it does not
Self-hosted model on your own hardware	the provider entirely; there is no provider	your own network boundary and your own patch discipline
Confidential computing at a cloud provider	the operator’s software and administrators as adversaries	the operator’s physical access, availability, side channels; and you are usually not the one verifying
Enterprise contract with a private network path	public internet exposure, and an unmanaged endpoint	the provider receiving and processing your plaintext

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

Destination	What it protects against	What it does not
Enterprise or API contract, public endpoint	training on your data, by default, under most business terms	retention for abuse monitoring, subprocessors, court orders, terms changes
Consumer or free tier	nothing you can rely on	everything above, plus training in some configurations

Restricted information belongs in row one. If that is not possible, the honest position is that it is an accepted risk, recorded as such, and not a protected trade secret. We would rather help you write that down than help you pretend otherwise.

SECTION 5

The contract decides more than the settings

This is the part that surprises people, and it is the cheapest thing on this list to get right.

- **Paying more does not change the contract class.** Anthropic's Pro and Max and OpenAI's Plus and Pro are consumer plans at every price tier. OpenAI's own naming says it: "business data" means Business, Enterprise and Edu. Business and enterprise agreements are a different product with different terms, not a bigger version of the same one.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

- **The free developer tier is the trap.** A Google AI Studio key with no billing attached falls under the Unpaid Services terms: the content may be used for training and may be read by human reviewers. Google's own terms tell you not to submit confidential information. A developer trying something out on a Friday is the most likely source of a disclosure in the whole organization.

- **Zero data retention is a gated product, not a switch.** It exists at several providers in several different shapes. Most require prior approval, every one of them carries a list of excluded features, and on the OpenAI API almost everything that makes an agent stateful, Conversations, Agents, Assistants and Threads among them, is marked not eligible. There is no switch there either, only a configuration checklist and an approval, and the `store` parameter that keeps your request defaults to on. Turning that parameter off is a request setting rather than a retention guarantee, and the next bullet has a provider saying exactly that about its own platform.

- **Abuse-monitoring retention can survive a zero-retention agreement.** This is not a commentator's inference; three providers say it on their own pages. Ask specifically whether abuse logging is inside or outside the addendum you are signing, and ask what you can check for yourself. Amazon Bedrock is the one platform we read that exposes a setting you can inspect: a per-account retention scope whose values run from `none` through `default` to `aws_review`, and whose own documentation warns that "setting `store=false` does not guarantee zero data retention". Elsewhere the only enforceable control is a policy you apply to your own account rather than a term in the contract.

- **The retention floor and the retention ceiling are different products.** An OpenAI Enterprise retention policy has a minimum of 90 days: an administrator cannot set it shorter. Anthropic's commercial API baseline is a 30-day maximum. Those are opposite instruments and they get quoted as if they were the same thing.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

- **Newest models can override it.** Anthropic's Covered Models policy, effective 2026-06-09, forces a retention window for its newest models on every platform that serves them. A zero-retention agreement signed in 2025 does not get you the 2026 models on the old terms.

- **A court can order retention regardless.** In May 2025, in the New York Times litigation, a federal court ordered OpenAI to preserve output logs "regardless of whether deletion was requested by users or mandated by privacy regulations". Enterprise and education agreements were carved out, and so were API customers who held a zero-retention agreement. The tier you had bought, and whether you had that agreement, decided whether you were inside the carve-out. The going-forward obligation was later released, in September 2025, which is the part usually left out: the lesson is not that the order still binds, it is that an order like it can arrive at any time and your contract decides where it lands.[4]

- **The same model can give opposite answers depending on who resells it.** Run Anthropic's models on Amazon Bedrock and retained content is not shared with Anthropic; AWS performs the review itself. Run the same models on Google Cloud and the customer must enable sharing that data with Anthropic for abuse monitoring in order to use them at all. The model is identical. The data path is not. This is the single clearest illustration in the paper that the destination is a contract, not a technology.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

- **Subprocessor notice is a term worth reading.** Microsoft's current Products and Services Data Protection Addendum gives at least six months' notice before a new subprocessor gets access to customer data, but only 30 days for subprocessors that support artificial intelligence functionality, with "the ability to disable use of that subprocessor until at least 6 months after notice".[5] Both halves matter, and the second half is the one the 30-day figure alone overstates past. Worth knowing too that Microsoft's own trust center page still states the six-month rule with no AI carve-out at all: the addendum is the governing document and it is newer. When a vendor's marketing page and its contract disagree, you are holding the contract.

- **Read "FedRAMP" as a listing, not an adjective.** We checked the live marketplace rather than vendor claims, and the picture is not what the marketing suggests. Of five well-known options, exactly one appears inside a certified service list. One vendor's own listing reads "Not yet certified" with zero authorizations. One hosted AI service that buyers widely assume is covered appears only under "Other Service(s)", never under Certified Services, and the two Certified Services lists it could have appeared on are empty, which is what makes the placement decisive rather than merely untidy. Two newer listings are pilots, one at Low impact with two authorizations and one at Moderate with three. If a certification matters to your procurement, look the product up by its listing identifier and read which list the service you are buying actually sits in.[6]

None of this requires a security team. It requires somebody to read the agreement and write down which tier each team is actually on. In our experience that exercise alone finds the biggest exposure in the building.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

SECTION 6

Step three: the network controls, and what each one actually buys

Tunnels and VPNs

Kind	The threat it was built for	What it does for this problem
Commercial VPN (the consumer brands)	a hostile coffee-shop network, your ISP watching, geolocation by IP	**nothing**
Corporate remote-access VPN	reaching internal systems, authenticating the user, forcing traffic through inspection	nothing about the provider; and split tunnelling quietly removes the inspection benefit
Site-to-site tunnel	confidentiality between two sites you control	nothing about a third-party service
Mesh overlay (the modern peer-to-peer ones)	reaching **your own** self-hosted systems without exposing them to the internet	genuinely useful, for that job only

A commercial VPN changes which IP address the AI provider writes in its log. The prompt, the attached document and the personal data inside it arrive exactly as they would have without it. Sold as protecting what you send to an AI service, a consumer VPN is security theatre. It is a good control for the threat it was designed for and the wrong control here.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

Two caveats worth saying out loud. Mesh overlays are the right tool for private access to your own inference server, and each one has a trust dependency worth naming: a coordination server, a relay, a certificate authority, or an account binding. And a VPN appliance is itself an attack surface with a long record of exploited vulnerabilities, so adding a concentrator can enlarge the problem you were trying to shrink.

Private endpoints into hosted AI

All three major clouds offer a private path to AI services: AWS PrivateLink, Azure Private Endpoint, and Google Private Service Connect. On Google the support is per-surface rather than blanket, across more than a dozen Vertex AI surfaces and several different access mechanisms, so do not write “Vertex AI supports PSC” into a design document. Check the specific API you will call, not the product family. What they genuinely remove is public internet transit, a publicly reachable endpoint, and dependence on an internet gateway. That is worth money for governance and for shrinking your exposed surface.

What none of them removes is that **the provider still receives and processes your plaintext.** Microsoft and Google both say in their own documentation that the traffic already stayed on their backbone, so the gain was never confidentiality.

A private endpoint presented to a board as “the data never leaves the company network” is a misstatement. It left. It left over a private path. And private is not the same as encrypted. AWS’s own words: “PrivateLink does not provide any encryption by default for data in transit.” Google’s are blunter still, and are the sentence to remember out of this whole section: both the public and the private path “offer SSL/TLS encryption. However, the private option offers much better performance” and is therefore recommended. That is a performance product, honestly described.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

Practical traps, since these are fiddly to configure: endpoint policies and private DNS on AWS; three private DNS zones, a mandatory custom subdomain on the resource, and a network-ACL bypass setting that can silently reopen a path you believe is closed on Azure; and a split between three different access products on Google. Per-endpoint hourly charges apply on all three. AWS and Azure add a data-processing charge per gigabyte and bill the hourly charge per availability zone. Five endpoint families across two zones is ten billed endpoint-hours, or roughly seventy-three dollars a month before a byte moves, and that is a floor rather than a quote. Google charges one cent an hour to reach its own APIs with no data charge and no zone multiplier, which is a genuine architectural difference rather than a pricing detail. Cite the API surface in your configuration rather than the vendor's brand name, because the brand names move.

Single-tenant hardware and confidential computing

Single tenancy buys you no co-tenants. It buys nothing against "the provider can read my data." DigitalOcean's bare-metal GPU page markets single-tenant privacy while noting that out-of-band management access is not provided to the customer, which is the whole argument in miniature: the management controller exists, you do not get it, the provider does.

Confidential computing is different, and it is the honest second choice. It genuinely removes the hypervisor, the host operating system and the provider's administrators as **software** adversaries. A provider who dumps your virtual machine's memory gets ciphertext. Do not understate that.

The products have names, and so do the exclusions: Intel TDX and AMD SEV-SNP on the processor side, NVIDIA confidential computing on the GPU. Read the published exclusions, because there are five of them and most are the vendors' own words:

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

1. **Availability is out of scope.** All three vendors say so. A hostile operator can simply never run your machine, or terminate it.

2. **Physical attacks on the memory bus are out of scope**, and the provider is by definition the party with physical access. AMD has declined mitigations in writing three times (bulletins AMD-SB-3024, -3040 and -3048, the last declining even to assign a CVE), plus a fourth for voltage fault injection. Intel states that TDX does not defend against replay of memory through physical attacks, **while its own published adversary model names a cloud-provider technician hooking into the DDR bus.** That is a structural contradiction at the center of the proposition.

3. **Side channels.** Be precise about who says this. AMD does publish the exclusion in its own security bulletin (AMD-SB-3035: it believes all the side-channel techniques discussed fall outside the SEV-SNP threat model). Intel publishes a chapter of mitigations instead and states no blanket exclusion, so there the out-of-scope characterization is the researchers', not the vendor's. Either way the newest research lands directly on AI. A 2026 paper, TDXRay, produced cache-line-granular memory access traces of unmodified confidential virtual machines and used them to infer user prompts by watching memory accesses during tokenization, recovering prompts from a single trace, **while fully complying with the platform's stated security model.**[7] The authors notified six major AI companies in November 2025. The fix requires new silicon.

4. **GPU memory is not encrypted.** NVIDIA's own words: the on-package memory is not encrypted, and the compute engines execute unencrypted code on unencrypted data resident in GPU memory. What is encrypted is the bus. On the current widely deployed generation the only multi-GPU mode leaves the GPU-to-GPU link unencrypted.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

5. **Firmware and microcode are inside the trusted base** by the vendors' own definitions, and both have had vulnerabilities whose stated impact is loss of exactly the protection being sold.

Two further points your procurement people will not hear from a vendor.

Vendor severity ratings systematically understate this class of problem. A complete bypass of Intel's named single-step defense, the Instruction-Count Single-Step Defense, scored 2.5 out of 10. A later defect in the same area, a predictable random-number seed, scored 3.3 on that same scale, and 2.0 on the newer version of it, which is worth knowing before anyone compares two scores without checking which scale each came from. The ratings are low because the prerequisite is a "privileged local attacker", which is **precisely the adversary the product exists to defeat.** A patching policy driven by severity scores will systematically deprioritise the entire category.

Ask who holds the deciding role in attestation. Attestation is the mechanism that proves the machine is running what it claims. On NVIDIA GPU confidential computing the party that decides is the NVIDIA Remote Attestation Service. On AWS Nitro Enclaves it is an AWS KMS key policy keyed to the enclave's measurements, and the failure mode there is quiet: a debug enclave emits all-zero measurements, so a key policy that does not pin the non-zero ones can be satisfied by a debug enclave. In Apple's published Private Cloud Compute design, the party that decides is the end user's own device. That last arrangement is the only one where the decider is not the provider or a service the provider runs, and it is the right yardstick to hold any vendor's attestation story against. NVIDIA's own sentence is the one to remember: attestation does not establish that approved code is free of vulnerabilities. It proves what is running, not that what is running is sound. Apple's guide is honest enough to say the matching thing about itself, that it is not possible to prove the published binaries were compiled from the published source. Treat Private Cloud Compute as the trust-model benchmark and the

design template, not as something you can buy: we found no Apple document offering it to third parties.

SECTION 7

Step four: self-hosting, done properly

Self-hosting is the only control in this paper that actually removes the provider, because there is no provider. It relocates the risk onto your boundary and your patching, and that trade is usually worth making. It is also the thing most commonly done badly, and the field evidence is not ambiguous.

The inference servers people actually install ship with very different defaults. Ollama and llama.cpp both bind to the local machine only, and Ollama has no authentication at all by design, which is the correct pairing, but Ollama's official container quickstart publishes the port and inverts it. vLLM **binds to all network interfaces by default** and its own documentation says in a heading that its API key does not protect every endpoint, that the key covers three path prefixes only, that other endpoints on the same server are unauthenticated including one exposing the same inference capability, and instructs readers: "Do not rely on `--api-key` alone to secure vLLM". Hugging Face's Text Generation Inference, still widely referenced in tutorials, **was archived in March 2026** and must not be specified in a new build.

Five independent scanning efforts over two years found internet-exposed Ollama instances: over 1,000, then roughly 10,000 with about one in four vulnerable, then 1,139 counted as vulnerable rather than merely exposed, then 12,269 with no authentication at all, then roughly 300,000. We print the series rather than a trend line because it is not one, and because the five efforts were not counting the same

thing. What they agree on is that the number is large and that nobody knows the denominator. One research team described an exposed instance as equivalent to exposing the container daemon socket.

So a self-hosting recommendation is incomplete unless it states all five of these:

1. The bind address, and that it is not all interfaces.
2. Authentication terminated at a reverse proxy that covers **every** path, not the engine's own partial key.
3. No published container ports on anything internet-facing, and no root in the container.
4. A patch cadence with a named owner, because these projects move fast and the vulnerabilities have been severe.
5. An external scan of your own address space, performed by you, on a schedule. Nobody has ever discovered an exposed inference server by re-reading their own configuration.

SECTION 8

Step five: one brokered way out

Default-deny outbound is the baseline, and then the question is what the allow list looks like.

- **IP allow lists break on content-delivery-fronted APIs.** Four well-known AI endpoints, resolved on 2026-10-24 with `dig` and `whois` from a single resolver: a method a reader can repeat in a minute rather than a permanent fact. `api.openai.com` resolves into Cloudflare's address space, where an allow list is either unmaintainable or means permitting a large fraction of the web. The other three resolve into their own operator's space, which sounds better than it is: `bedrock-runtime.us-east-1.amazonaws.com` and `generativelanguage.googleapis.com` sit inside Amazon's and Google's own enormous published ranges, so an allow list built on them permits most of a cloud. Exactly one of the four, `api.anthropic.com`, sits in a narrow published range that the vendor says "will not change without notice". Which vendor you picked decides whether this control is even available to you.

- **Name-based allow listing is the workable form, and it is not an anti-exfiltration control.** The proof is in the vendors' own setup documentation: Claude Code's enterprise network page lists the hosts to allow and introduces them with "among them", so the list is explicitly partial. It names at least twenty-five, including `github.com`, `registry.npmjs.org`, `raw.githubusercontent.com`, `storage.googleapis.com`, two telemetry intake hosts, a font service and five JavaScript content-delivery networks. Every one of those is a destination somebody else can write to, and the vendor does not publish a total.

- **Inspecting TLS gives you content visibility and hands you a liability.** The proxy now holds decrypted personal data and becomes a target in its own right. If you do it, somebody has to own that archive on purpose.

- **The pattern that actually works is a single brokered exit:** one gateway that holds the credentials, applies redaction, logs the traffic, enforces per-team policy, and can be switched off. One place to audit beats every workstation talking to the internet directly. Accept what it does not fix, though, because it is the honest limit of this whole section: the approved, allow-listed, logged, authenticated path to the model API is itself a full-bandwidth channel out of your network. OpenAI's own paper on this problem reports that its hand-curated allow list "covers only approximately ten percent of URLs that our users were actually visiting", which is why allow lists get widened until they stop meaning anything, and it describes its own mitigations as "compensating controls while model-intrinsic robustness to web-based prompt injection is still evolving". And the Model Context Protocol, the tool-connection standard everyone is adopting, says in its own specification that the client acts as a proxy and bypasses network perimeter controls, and tells implementers to use egress proxies. Take that as the spec authors agreeing with us.

- **DNS is an egress channel and it is usually unlogged.** Encrypted DNS bypasses your resolver policy unless you block it, and DNS queries do not appear in the standard cloud network flow logs, so an organization that believes it has complete network logging does not.

- **Two standards changes are coming for anyone planning this.** Encrypted Client Hello became a Standards Track standard in March 2026 as RFC 9849, and it hides the server name your own egress firewall inspects to decide what to allow. AWS Network Firewall already documents the exception: it supports TLS 1.1 through 1.3 "with the exception of encrypted client hello (ECH) and encrypted SNI (ESNI)", and AWS's own guidance says the firewall resets the connection when ECH stops it reading the name. Separately, the certificate lifetime is being cut from 398 days to 47 by 2029, which ends static certificate pinning as a maintainable control. If your design depends on either name inspection or pinning, design for the replacement now.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

SECTION 9

Step six: rules for AI agents

An agent is different from a chat window in one way that matters: it acts. It reads your files, calls your systems and fetches web pages, and anything it reads can try to instruct it.

The rules below are the ones we would write into a policy. They are ordered by how much they buy.

1. **Classify first, then route.** The agent that handles restricted material and the agent that drafts marketing copy are not the same agent and do not share credentials.
2. **Least privilege, and mean it.** Read-only by default. Scoped credentials per agent, never a shared administrative one. No direct database access; go through a gateway that exposes named operations, so that the real table and column names never cross the wire. The reason is not elegance: a leaked query cannot describe a schema it never contained, and an agent that never held a schema cannot be talked into reciting one.
3. **Allow-list the tools, not just the domains.** See the next section for why.
4. **Approval gates on anything irreversible or outward-facing.** Deleting, paying, sending, publishing, deploying, changing credentials. A human presses the button.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

5. **Deterministic redaction before the request leaves**, and do not trust it. The argument here is not ours, it is theirs. Presidio: “there is no guarantee that Presidio will find all sensitive information.” Google Cloud Sensitive Data Protection: “Built-in infoType detectors are not a perfectly accurate detection method.” Microsoft’s own legal transparency note for Azure AI Language PII detection carries the section heading “Avoid high-risk automatic redaction or information classification scenarios.” Amazon Macie: “Any single object sample isn’t guaranteed to be conclusive.” Three vendors and one open-source project, each disclaiming completeness in writing. A control its own maker will not warrant cannot be the last line of defense. Named failure modes include recall gaps with demographic bias, the health-privacy safe harbour’s catch-all for “[a]ny other unique identifying number, characteristic, or code”, which no tool can implement, and re-identification from combinations of innocuous fields.

6. **An LLM must never be the component that decides whether data is safe to send.** This is the one non-negotiable on the list. The decision has to be made by rules you can read, in code you control, over categories you defined.

7. **Log every outbound payload, and then protect the log.** This control carries its own paradox: the audit trail now contains the sensitive data, in one place, structured for convenient reading. DeepSeek left exactly such a log database exposed in January 2025: an unauthenticated ClickHouse instance, more than a million rows of logs, API keys included, found by the security firm Wiz rather than by DeepSeek. Nobody had to be clever. If you build the log, it inherits the classification of the most sensitive thing in it.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

SECTION 10

Prompt injection: plan for it, because nobody has solved it

If an AI agent reads a document, a web page, an email or a code comment, whatever is written there can attempt to give it instructions. This is called indirect prompt injection and it is the central unsolved problem in the field.

We are not editorializing. The consensus is unanimous and citable. OWASP's own guidance says in terms that no reliable prevention mechanism exists today. The UK National Cyber Security Centre says such attacks may never be totally mitigated and the best available outcome is reducing likelihood or impact, in a post titled "Prompt injection is not SQL injection". OpenAI's chief information security officer calls it a frontier, unsolved security problem. And a 2025 academic paper took twelve recent published defenses, the majority of which had reported near-zero attack success, and got above 90% on most of them.[8] **That last result is the one to quote at any vendor who tells you they filter it.**

Two pieces of honesty now, in opposite directions, because the scare stories and the reassurances are each half the picture.

First, on indirect prompt injection specifically: **every exfiltration case we could document is demonstrated, reported and patched, and none of it has yet been observed in the wild.** The relevant vendor severity records carry exploit maturity "unproven", and the one entry with national vulnerability enrichment data records exploitation "none". We say "not yet observed" rather than "does not happen", because that is a null result against the sources we read, not a property of the attack. One lead that may change it is unverified and we are not citing it as fact.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

Second, and this is the correction to the comfortable reading of that paragraph: **real adversaries are already stealing real data through AI integrations. They are simply not bothering with prompt injection to do it.** The largest confirmed case we found, Salesloft Drift in August 2025, involved no injected prompt at all. Drift is an AI chat agent, and its Salesforce integration held OAuth and refresh tokens; attackers stole the tokens and queried customer Salesforce instances directly, **bypassing multi-factor authentication**, running automated exports and deleting the query job records behind them. What they were after was other credentials sitting in CRM records: cloud access keys, passwords, warehouse tokens, VPN and single-sign-on URLs. Google's Mandiant attributed it to the actor it tracks as UNC6395 and said more than seven hundred organizations may have been affected. The AI was never the leak path; its integration privileges were. Separately, a malicious npm package named `postmark-mcp` silently copied outbound email to an attacker across an estimated three hundred organizations. And Asana's experimental connector exposed data across roughly a thousand customers through a plain tenant-isolation bug which, in the words of the security firm that wrote it up, was "not caused by an external hack".

Hold those two paragraphs together, because the design conclusion arrives from both ends at once. The interesting question was never whether the model can be tricked. It is what the agent's credentials can reach once anybody at all gets hold of them.

The mechanical attack worth understanding is boring and procedural: a model emits a link to an image at an attacker's address with your data in the query string, and the client fetches it with no click required. A content security policy constraining where images may load from is the right shape of fix, and it is not a solution. Three of the documented defeats went straight through one: Google Apps Script on `script.google.com`, a `*.cloudfront.net` wildcard, and, in the Microsoft 365 Copilot case CVE-2025-32711, "abusing a Microsoft Teams proxy allowed by the content security policy". **An allow-list entry that can be made to log a request is not an allow-list entry.** Note also what none of this is:

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

none of it is enforceable by your firewall. These are fixes only the vendor can ship, which is a reason to ask a vendor what it has shipped.

The finding to carry into your design is what the documented cases have in common. In several of them, including Salesforce ForcedLeak, where a domain on Salesforce's own content-security-policy allow list had expired and the researchers bought it for five dollars, the data left through a destination the defender had **deliberately allowed**, using a credential the defender had **deliberately issued.** Three of our research rounds reached that conclusion independently from different evidence.

So: **the control that matters is what the agent can reach, not which domains it may talk to.** An egress allow list is necessary and nowhere near sufficient.

SECTION 11

People, accidents and the limits of what anyone knows

Most disclosure is not an attack. It is somebody pasting a contract into the wrong window.

We want to be careful here, because this is where the numbers in circulation are worst. **We could not find a government or academic prevalence survey of unsanctioned AI use**, and we ran twenty-one named searches looking for one; that is a located null result, not a proof that none exists. What the market offers instead is vendor telemetry, and five well-known vendor figures disagree with each other because their denominators are different things: users, requests, access events, session minutes, and applications. We hold those five side by side

as a disagreement and **we will not average them**, because averaging incommensurable denominators produces a number that means nothing and sounds authoritative.

The one figure with a fully published independent sample comes from IBM's annual Cost of a Data Breach report, the research conducted by the Ponemon Institute: incidents involving unsanctioned AI more than doubled year on year, to 43% from 20%, across 3,558 interviews at 602 organizations. Use that one, and use it as a direction of travel rather than as your own exposure.

What we will say with confidence is the mechanism, and with humility the magnitude.

- The widely repeated 2023 origin story is told wrong in both directions. Samsung's restriction is independently confirmed: it told CNBC it was "temporarily restricting the use of generative AI through the company's personal computers", which is narrower than the outright "ban" of the retellings. What rests on a single Korean business magazine, Economist Korea and not The Economist of London, is the colorful part: three incidents in twenty days, with Samsung itself saying it was "difficult to confirm" the internal matter. And no source shows Samsung's code reaching anybody other than OpenAI, so "Samsung's source code leaked" overstates it; what is documented is confidential source code disclosed to a third-party service outside the company's control, twenty days after the tool was allowed. The second famous case, Amazon in January 2023, was a lawyer warning staff not to paste Amazon code into ChatGPT, together with a claim that output had already closely matched internal material. A warning and a claim, not a documented breach. Cite both as what they are, and watch which half of a story your source is actually carrying. After those two the public record goes quiet, and that is itself a finding: we could verify no named post-2023 enterprise case of accidental disclosure to a hosted AI service. The 2025 and 2026 incident literature is vulnerability research, not employee pastes at named companies.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

- **The same editor and the same keystrokes can produce opposite data postures depending on the license tier.** GitHub's Terms of Service define an AI Feature's Input as prompts plus "attachments, **code in your workspace**, and conversation history", and grant GitHub "a license to collect and use your Inputs and Outputs to develop, train and improve artificial intelligence" models unless the developer opts out in account settings or the seat is governed by another agreement. Those training provisions "apply only to individual licenses"; a GitHub Copilot Business or Enterprise seat bought directly from GitHub carries the opposite commitment in writing, that "GitHub will not use Inputs or Outputs to train generative AI models".[9] So a developer on a personal subscription is licensing their workspace code for training by default, and the developer at the next desk on a Business seat is not, in the same editor. Cursor is confusing in a different way: its privacy policy says "We do not use Inputs or Suggestions to train our models" outside three narrow exceptions, while its own privacy overview says that with Privacy Mode off it may use stored codebase data and prompts to train its models. Neither page states a retention period.[10] Your developers are not being careless. **Ask which license each seat is actually on, and read that tier's terms rather than the vendor's front page.**

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

- **A coding assistant that saves a transcript of the work creates a second copy of everything discussed in it, automatically, on local disk, and this is the item most often missing from a data inventory.** It is documented behavior, not a vulnerability: Anthropic publishes that Claude Code stores session transcripts locally in plaintext under the user's home directory, 30 days by default, and the vendor's own pages do not agree on that figure, so treat it as what the page says rather than as a settled fact.[11] The point is what the file is. A transcript of a design discussion is a record of the design, which makes the transcript store intellectual property in its own right, sitting outside the repository, outside source control and outside whatever retention policy covers the code it discusses. Because a tool creates it rather than a person, three questions usually have no answer: where it lives, how long it stays, and what classification it inherits. Each has real trade-offs, and a transcript is also the thing a developer most wants to keep. **Ask what your tools write to disk, then decide those three deliberately, in that order.** An inventory that covers the repository and not the transcript store covers the asset and not its shadow.

- **The same class of tool ships commands that send the conversation off the machine on purpose.** Anthropic documents that Claude Code's feedback, bug and share commands each send "a copy of your conversation history including code" to the vendor, retained five years, and that one of them can optionally open an issue in a public repository.[11] A developer trying to be helpful about a bug can move client code across that line with one command, and no network control will stop it, because it is the product working as designed. We cite this product because its documentation states the behavior plainly enough to quote, not because it is unusual; the correct assumption is that your assistant has an equivalent. **Find those commands in whatever tools your developers run, and put them in the policy rather than leaving them to judgment.**

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

SECTION 12

Where the law actually stands

Short version: the exposure is real, the case law is thin, and anyone telling you it is settled is selling something.

Trade secrets. The federal definition, 18 U.S.C. 1839(3)(A), requires the owner to take "reasonable measures to keep such information secret". That is a definitional element, not a defense: fail it and there is no asset to protect, not merely a weaker case. The engine behind this is *Ruckelshaus v. Monsanto Co.*, 467 U.S. 986, 1002 (1984), holding that disclosure to someone under no obligation to protect confidentiality extinguishes the property right. It is not AI law; it does not need to be.

One decided case is on point, and we will describe it accurately. In January 2026, in *Trinidad v. OpenAI*, a federal district court held that it could not conclude the plaintiff's "protocols and frameworks" were trade secrets, because she admitted developing them through ChatGPT, which was voluntary disclosure to a party under no duty of confidence. The court added that this held **whether or not** her consent was enforceable as a contractual matter.[12] That reasoning is directly relevant. It is also unpublished, non-precedential, brought by a self-represented plaintiff, decided on the pleadings, and not reasoned in AI-specific terms. Anyone citing it as settled law is overreaching.

The proposition everyone is afraid of is untested. No court has held that failing to prohibit or control employee AI use is itself a failure of reasonable measures. We searched for one, and the population is worth stating plainly: the federal opinion and docket corpus through a public court-records interface and its document archive, three query formulations plus variants, law firm client alerts, and a 2026

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

continuing-education deck. That is a null result against a named population, not a proof of absence, and it would not catch an unreported state court ruling. The cleanest evidence that it is untested: a 2025 case that pleaded prompt injection as misappropriation **dropped the trade secret claim on amendment.**

And an enterprise contract is one exhibit, not the answer. A 2026 decision held that non-disclosure agreements are not dispositive of whether a plaintiff made reasonable efforts. And abuse-monitoring retention can persist even under a zero-retention addendum, which is not a commentator's reading of the contract but a thing three providers state on their own pages. Signing the right paper is necessary and is not the whole of "reasonable measures".

Privilege does not survive the consumer tier. In February 2026, in *United States v. Heppner*, Judge Rakoff of the Southern District of New York held that a defendant's conversations with a consumer AI assistant were neither privileged nor work product, treating the question as one of first impression nationwide, while preserving the established carve-out for a tool used at the direction of counsel.[13] One note on how we read it: that opinion is a scanned image. We verified its caption, docket number and filing date from the primary document, and the holding itself is quoted from a law firm's analysis with pin cites rather than read off the page.

The professional guidance is ahead of the case law, and it is the most useful thing in this section because it tells you what diligence looks like. The California State Bar's 2026 "Practical Guidance for the Use of Generative Artificial Intelligence in the Practice of Law", revised at the California Supreme Court's request to address agentic AI, says that reasonable efforts require more than reliance on generalised marketing assurances, and that the degree of diligence and supervision must correspond to the level of system access and autonomy the tool has. Both halves of that are a usable standard for any organization, not just a law firm.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

Sector rules attach to the vendor relationship, not the technology. Under HIPAA, a model vendor processing protected health information on a covered entity's behalf is a business associate and needs an agreement. The Gramm-Leach-Bliley safeguards rules require a written programme and service-provider oversight. FERPA's school-official exception requires direct control over the vendor's use. CJIS imposes personnel screening on provider staff, and there is no AI-specific pathway in that policy and no certification you can ask a model vendor to produce. And under ITAR and the EAR, sending technical data to a foreign person is an export, including a foreign national performing support at the vendor. In every one of these regimes the obligation lands on the contract and the record of it.

The export rules also contain the sharpest statement of this paper's central point, and it is in the regulation rather than in anyone's marketing. Both export regimes treat properly encrypted technical data as not an export, and both condition that on the same thing: the means of decryption are not provided to any third party. That is the legal test and it is the engineering test. Transport encryption to a provider who decrypts at the other end does not pass it, because you handed over the means.

One accuracy note, because it is the kind of thing that gets repeated wrongly: of the six frameworks vendors cite at you, only two are certifiable, ISO/IEC 42001 for an AI management system and ISO/IEC 27001 for information security. Both carry accredited third-party certification; we read AWS's own 42001 certificate, issued by a body accredited by the ANSI National Accreditation Board. SOC 2 is an **attestation**, which is an auditor's opinion on controls the vendor itself selected. NIST's AI Risk Management Framework gives you a shared vocabulary and nothing to fail. Ask for the SOC 2 report, then read its scope and its exceptions.

This section is general information and not legal advice. Take the specifics to your counsel.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

SECTION 13

The short version

If you do nothing else from this paper:

1. Write down the four categories and which destination each one may go to. One page.
2. Find out which tier each team is actually on, including anybody using a free developer key. This is the single highest-yield hour in the project.
3. Put restricted material on hardware you own, or record in writing that you have accepted the risk and do not claim it as a trade secret.
4. Stop paying for controls that address a different threat. A consumer VPN does nothing here. A private endpoint buys governance, not confidentiality. Single tenancy buys neither.
5. If you self-host, state the bind address, terminate authentication at a proxy covering every path, publish no container ports, name a patching owner, and scan your own address space.
6. Route everything through one brokered exit that holds the credentials and the log.
7. Give agents read-only, scoped, named access, approval gates on anything irreversible, and never let the model decide what is safe to send.
8. Assume prompt injection works, and design so that a successful injection cannot reach anywhere useful.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

9. Ask what your own tools write to disk, and what their feedback commands send.

Items 1, 2 and 9 cost nothing but attention and find most of the exposure.

Who We Are

White Rock Future Tech LLC provides AI-related services for small to large businesses. We also offer CIO/CTO services on a part-time and consulting basis. Our company uses AI in our daily work, which is kept in house on private networks when dealing with client materials and IP. We check each claim from our AI agents against their citations as well as checking for supporting information.

AI is a potentially game-changing tool for business today. We believe that the use of AI, in the correct hands and with appropriate guardrails, boundaries and rules will be as transformative as the introduction of the Personal Computer (PC) was back in the late ’70s and early ’80s.

Glossary

The terms used in this paper are explained in plain language on a separate page: Glossary for VPNs and TLS Protect the Pipe, Not What AI Providers Keep.

(352) 614-8226
info@wrfuturetech.com
www.wrfuturetech.com

Sources

The 13 sources behind this paper, with links, are listed on a separate page: Sources for VPNs and TLS Protect the Pipe, Not What AI Providers Keep.

This page is general information, not legal advice.