

A SUPERWISE® LEARN KIT · OCTOBER 2026

# What can be audited and guaranteed: vendor claims, decoded

Fourteen claims AI vendors make, what each can honestly mean, the proof to ask for, and the answer that should end the conversation. Take it into the next vendor call.

# Turn every claim into a promise you can check

A claim worth relying on names four things: the exact product and feature, the boundary where the control sits, the exceptions, and the evidence you can inspect. A bare adjective or a round percentage names none of them.

- 01 **Collect the claims.** Pull every AI statement from the proposal, the website, the security questionnaire, and the contract into one list.
- 02 **Find each one in part 2.** The table gives the honest reading, the proof to request, and the red flag. Rows are grouped by what the claim is about: your data, the paperwork, and the model's behavior.
- 03 **Ask the vendor to complete the sentence below.** Use the worksheet in part 3 during the call. A claim the vendor cannot complete is marketing, and it goes in your notes that way.
- 04 **Get the answer in writing.** The signed order form, data processing agreement, business associate agreement, and the vendor's published exception lists decide what is true for you.

## The evidence sentence

For [exact product, plan, and feature], when [defined input and configuration] crosses [named control], the system will [observable result], except [listed cases], and you can verify it through [log, test, report, or contract clause] kept for [period].

### PLAIN ENGLISH • GUARANTEE

A promise about a result that the vendor's own software enforces every time, on traffic that reaches it, and that you can check afterwards. Blocking a request at a gateway can be guaranteed. A model's accuracy is a measurement: a rate observed on a test, which changes with the data and the model version.

**Vendor terms move fast.** Several pages cited in this kit changed in the week it was written. Every vendor fact carries the date it was read, and the numbers in gray point to the sources at the end. Check the current page before you rely on one.

**Not legal advice.** This kit summarizes public vendor documentation and published standards. Confirm contract decisions with counsel.

# Fourteen claims, decoded

Grouped by what the claim is about. Vendor facts were read on the vendor’s own page on October 8, 2026, unless a row says otherwise.

WHERE YOUR DATA GOES

| THE CLAIM                      | WHAT IT CAN HONESTLY MEAN                                                                                                                                                                                                                                                                                                                                                                                        | ASK FOR THIS AS PROOF                                                                                                                                                           | RED FLAG                                                                                      |
|--------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| "Zero data retention"          | Named products or endpoints keep no copy of prompts and outputs after processing, under an approved agreement. Anthropic applies it only to eligible API products and Claude Code for Enterprise, and still keeps safety classifier results [1]. Its designated covered models keep prompts and outputs for 30 days [2]. OpenAI requires prior approval, and some API endpoints still store data with it on [3]. | The contract clause. The list of covered products, endpoints, and models. What is still kept: safety records, files, batch jobs, feedback, and metadata.                        | It is described as covering "everything," including the consumer app, with no exception list. |
| "We do not train on your data" | Customer content from a named business product stays out of model training by default. OpenAI says so for its business products, unless you opt in [4]. Anthropic’s commercial terms make exceptions for feedback you report and programs you choose to join [5].                                                                                                                                                | Which content the promise covers: prompts, outputs, files, feedback, and flagged content. Whether an admin can stop users opting in. How long the vendor keeps the data anyway. | The promise is offered as proof of zero retention or of no human review. It covers neither.   |
| "Your data stays in region"    | Listed content is stored in a chosen region, and sometimes processed there. OpenAI stores ChatGPT Enterprise content at rest in 10 regions and offers in-region GPU inference in three. Authentication, routing, analytics, workspace metadata, and logins can sit elsewhere, and external apps, MCP servers, and web search follow the provider’s own terms [6].                                                | A split answer for storage, inference, routing, metadata, backups, support access, and connected apps. Which of those the contract pins to the region.                          | One region named for "your data," with no list of what is in scope.                           |
| "Encrypted end to end"         | Data is encrypted in transit and at rest, sometimes with a key you control. The model still has to read the content to work on it. Google states that Gemini in Workspace cannot act on email and files encrypted with customer-controlled keys [7].                                                                                                                                                             | Where the data is decrypted, who can use the key, and whether logs, backups, and the AI features themselves are covered.                                                        | "End to end" used for a service whose servers read your prompt to answer it.                  |

## THE PAPERWORK AND THE RECORDS

| THE CLAIM                                 | WHAT IT CAN HONESTLY MEAN                                                                                                                                                                                                                                                                                                                                 | ASK FOR THIS AS PROOF                                                                                                                                                                  | RED FLAG                                                                                   |
|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| <b>"HIPAA compliant"</b>                  | A named service can support your HIPAA obligations once a business associate agreement (BAA) is signed and you use only eligible features in the required configuration. HHS requires satisfactory assurances in a written business associate contract [8]. OpenAI and Anthropic offer a BAA only for named plans and features [9, 10]. Read Oct 7, 2026. | A BAA for this exact plan. The excluded features, connectors, and subprocessors. The configuration guide the BAA depends on.                                                           | A badge or the word "certified" in place of a signed BAA.                                  |
| <b>"Enterprise-grade security"</b>        | The vendor may offer single sign-on, access controls, encryption, an incident process, and independently examined controls. A SOC 2 report examines a service organization's controls relevant to security, availability, processing integrity, confidentiality, or privacy [11], for the system and period it describes.                                 | The SOC 2 report under NDA, its period, and the systems it covers. A letter covering the time since the period ended. A penetration test summary. A shared-responsibility matrix.      | No report is available, or the report covers a different product or company.               |
| <b>"Complete audit trail"</b>             | Named user and admin events are recorded, sometimes with prompts and outputs. OpenAI's Compliance Logs Platform is for Enterprise and Edu, append-only, and keeps logs for 30 days [12]. Anthropic's Enterprise audit export covers 180 days and leaves out chat and project content [13].                                                                | One sample event from each product surface you use. Whether it holds user, time, input, output, sources, tool calls, and policy decisions. Who can delete it, and how long it is kept. | "Everything is logged," with no sample event and no retention period.                      |
| <b>"The AI respects your permissions"</b> | The assistant retrieves content with the signed-in user's existing access. Microsoft states Copilot can only summarize or reference content the user is authorized to access, and points admins to oversharing controls [14]. Files shared too widely become easier to find.                                                                              | Which identity is used for retrieval and for actions. A test with a deliberately overshared file and a restricted one.                                                                 | Permissions offered as the whole answer, with no plan for files already shared too widely. |

## WHAT THE MODEL DOES

| THE CLAIM                                        | WHAT IT CAN HONESTLY MEAN                                                                                                                                                                                                                             | ASK FOR THIS AS PROOF                                                                                                                                                             | RED FLAG                                                                           |
|--------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| <b>"100% PII detection"</b>                      | A detector scored 100% on one test set, for a listed set of data types, languages, and formats. The open-source Presidio project states there is "no guarantee" it will find all sensitive information, and recommends additional protections [15].   | The test set, the data types, the languages, and the count of missed items. A run on a sample of your own documents.                                                              | The vendor will not run your sample, or cannot say what the test set was.          |
| <b>"Hallucination-free"</b>                      | A narrow workflow that quotes only retrieved passages, refuses outside its subject, or meets a measured error rate on a defined test. NIST names confabulation, confidently stated false content, as a generative AI risk to measure and manage [16]. | The definition of an error, the test set, the model version, and the observed error rate. Who checks high-stakes answers.                                                         | No error rate at all. A citation offered as proof the sentence around it is right. |
| <b>"99% accurate"</b>                            | One metric reached 99% on one dataset, on a stated date and model version. NIST calls for documented test sets, metrics, and limitations, assessed in the conditions of use [17].                                                                     | Accuracy of what, against whose labels, on what data. Precision and recall, the severity of the errors, and results on a sample like yours.                                       | The 1% is never described, or the test data looks nothing like your work.          |
| <b>"Guaranteed output" or "deterministic AI"</b> | A software layer after the model enforces a format, required fields, an allowlist, a refusal, or a block. Fixed settings can make repeat answers more stable. NIST treats validity and reliability as assessed under deployment conditions [17].      | Whether the guarantee is about format, action, or truth. The component that enforces it. What changes when the model or the source documents change.                              | A fixed "temperature" or seed offered as proof that answers are correct.           |
| <b>"Protected from prompt injection"</b>         | The system reduces specific attack paths with input isolation, least privilege, filtering, confirmations, and monitoring. OWASP states it is unclear whether fool-proof prevention exists [18].                                                       | The direct and indirect attacks tested, including text hidden in documents, web pages, and email. What the model is allowed to do. Which actions need approval outside the model. | "Blocks 100% of attacks," or no test results to show.                              |
| <b>"Human in the loop"</b>                       | A named person reviews a draft or approves a consequential action before it runs. NIST's AI Risk Management Framework calls for documented roles and human oversight processes suited to the context [17].                                            | Which actions pause, who can approve, what evidence they see, what happens on a rejection or a timeout, and an exported approval record.                                          | An approve button with no time, information, or authority behind it.               |

# What a vendor can promise, and what it cannot

### Promises with a defensible scope

- A request that crosses a named gateway is blocked by a fixed rule.
- A validator rejects output in the wrong format.
- An account without a permission is refused.
- A named event is written to an append-only log and kept for a stated period.
- Listed content is stored in a contracted region.
- A named product keeps customer content out of training, under contract.

### Promises nobody can keep

- A general-purpose model never states anything false.
- A detector finds every piece of sensitive data in every language and format.
- A model resists every future prompt-injection technique.
- A citation makes the answer around it correct.
- A certification covers an unnamed product, region, or future setup.
- Review works when the reviewer lacks time, context, or authority.

Every promise in the left column still needs a coverage statement. A guarantee about a gateway covers the traffic that reaches the gateway, and staff on a personal account in a browser are outside it. A residency promise about stored prompts leaves out the connectors. A log covers the product surfaces it ingests.

## Worksheet for the call

One row per claim. Leave a blank where the vendor had no answer, and follow up on the blanks in writing.

| PART OF THE SENTENCE       | CLAIM 1 | CLAIM 2 | CLAIM 3 |
|----------------------------|---------|---------|---------|
| The claim, as stated       |         |         |         |
| Product, plan, and feature |         |         |         |
| Named control              |         |         |         |
| Observable result          |         |         |         |
| Listed exceptions          |         |         |         |
| How we verify it           |         |         |         |
| Kept for                   |         |         |         |
| Where it is in writing     |         |         |         |

Vendor \_\_\_\_\_ Call date \_\_\_\_\_ Taken by \_\_\_\_\_

# SUPERWISE® Sentinel, held to this kit’s standard

Sentinel is an AI gateway that runs on the PC. Each claim below is quoted from [superwise.ai/sentinel/](https://superwise.ai/sentinel/) as read on October 8, 2026 [19], with its scope and the way to check it.

| THE CLAIM, AS PUBLISHED                                                                       | SCOPE                                                                                                                                                                                                                                                                                      | HOW TO CHECK IT                                                                                     |
|-----------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| Sensitive values are “replaced before the request leaves”, with placeholders such as <EMAIL>. | The published list: emails, phone numbers, account and card numbers, government IDs, API keys and access tokens, and dates of birth and medical record numbers when labeled. Names, street addresses, and passwords are absent from it. It applies to requests that pass through Sentinel. | Type a made-up client record into each approved tool and read what was replaced.                    |
| “Every inspected interaction lands in the Activity Log.”                                      | The log records what was checked, what was redacted, and when, for interactions Sentinel inspects. Traffic that never reaches Sentinel leaves no entry. The page itself calls the log “a record of what was inspected”.                                                                    | Match the log entries against the test prompts you sent.                                            |
| “It runs on the machine or in your own cloud, and nothing is routed through SUPERWISE.”       | Prompts go from the PC to the AI provider you already use.                                                                                                                                                                                                                                 | Have your IT lead watch the PC’s outbound connections while a test prompt is sent.                  |
| Where it works.                                                                               | Windows 10 and 11. Claude Code. <a href="https://claude.ai">claude.ai</a> and <a href="https://chatgpt.com">chatgpt.com</a> through a browser extension in preview for Chrome and Edge on Windows. Gemini, Copilot, and other chat sites are listed as in development.                     | Hold your approved tool list against the supported list. Anything outside it needs another control. |

## What the page does not claim

- **No certification.** Sentinel makes no tool HIPAA compliant, supplies no business associate agreement, and carries no certification claim of its own.
- **No discovery.** Sentinel sees the traffic sent through it. A tool on another device, or a personal account it does not cover, stays outside its record.
- **No detection rate.** The published list says which data types are replaced. Run your own sample to see how it handles your documents.

**Trying it.** Free for 30 days with no card, then \$10, \$25, or \$50 a month per deployment, as listed on [superwise.ai/pricing/](https://superwise.ai/pricing/) on October 8, 2026 [19]. A live demo is at [superwise.ai/sentinel/demo/](https://superwise.ai/sentinel/demo/).

# Sources

---

Vendor pages were read on the date shown. Several show only a relative update date; the one shown on the day is recorded with each.

- 1 Anthropic Privacy Center, *I have a zero data retention agreement with Anthropic. What products does it apply to?* Marked "Updated this week," read October 8, 2026. [privacy.claude.com/en/articles/8956058](https://privacy.claude.com/en/articles/8956058)
- 2 Anthropic Privacy Center, *Data retention practices for covered models*, dated September 5, 2026, read October 8, 2026. [privacy.claude.com/en/articles/15425996](https://privacy.claude.com/en/articles/15425996)
- 3 OpenAI, *Data controls in the OpenAI platform*. Zero Data Retention approval, abuse monitoring retention, and endpoint eligibility. Read October 8, 2026. [developers.openai.com/api/docs/guides/your-data](https://developers.openai.com/api/docs/guides/your-data)
- 4 OpenAI, *Enterprise privacy at OpenAI*. Read October 8, 2026. [openai.com/enterprise-privacy/](https://openai.com/enterprise-privacy/)
- 5 Anthropic Privacy Center, *Is my data used for model training?* Updated August 18, 2026, read October 8, 2026. [privacy.claude.com/en/articles/7996868](https://privacy.claude.com/en/articles/7996868)
- 6 OpenAI Help Center, *Data residency and inference residency for ChatGPT*. Marked "Updated 4 days ago," read October 8, 2026. [help.openai.com/en/articles/9903489](https://help.openai.com/en/articles/9903489)
- 7 Google Workspace Admin Help, *Gemini for Google Workspace FAQ*. Read October 8, 2026. [knowledge.workspace.google.com/admin/gemini/gemini-for-google-workspace-faq](https://knowledge.workspace.google.com/admin/gemini/gemini-for-google-workspace-faq)
- 8 HHS Office for Civil Rights, *Business Associates guidance*. Read October 6, 2026. [hhs.gov/hipaa/for-professionals/privacy/guidance/business-associates/](https://hhs.gov/hipaa/for-professionals/privacy/guidance/business-associates/)
- 9 OpenAI Help Center, *HIPAA eligible products and functionality*. Read October 7, 2026. [help.openai.com/en/articles/20001069](https://help.openai.com/en/articles/20001069)
- 10 Anthropic Privacy Center, *Business Associate Agreements (BAA) for Commercial Customers*. Read October 7, 2026. [privacy.claude.com/en/articles/8114513](https://privacy.claude.com/en/articles/8114513)
- 11 AICPA & CIMA, *System and Organization Controls: SOC suite of services*, and its publication on SOC 2 examinations of controls relevant to security, availability, processing integrity, confidentiality, or privacy. Read October 8, 2026. The landing page names the five categories only; the report period and system description are features of the report itself. [aicpa-cima.com/resources/landing/system-and-organization-controls-soc-suite-of-services](https://aicpa-cima.com/resources/landing/system-and-organization-controls-soc-suite-of-services)
- 12 OpenAI Help Center, *OpenAI Compliance Platform for Enterprise and Edu customers*. Marked "Updated 11 hours ago," read October 8, 2026. [help.openai.com/en/articles/9261474](https://help.openai.com/en/articles/9261474)
- 13 Claude Help Center, *Access audit logs*. Updated June 15, 2026, read October 8, 2026. [support.claude.com/en/articles/9970975](https://support.claude.com/en/articles/9970975)
- 14 Microsoft Learn, *Microsoft Copilot data protection architecture*. Updated August 18, 2026, read October 8, 2026. [learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-architecture-data-protection-auditing](https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-architecture-data-protection-auditing)
- 15 Presidio documentation, home page, on the limits of automated detection. Moved from [microsoft.github.io/presidio](https://microsoft.github.io/presidio); read October 8, 2026. [presidio.dataprivacystack.org](https://presidio.dataprivacystack.org)
- 16 National Institute of Standards and Technology (July 26, 2024), *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1. [nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf](https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf)
- 17 National Institute of Standards and Technology (January 26, 2023), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1. [nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf](https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf)
- 18 OWASP GenAI Security Project, *LLM01:2025 Prompt Injection*. Read October 8, 2026. [genai.owasp.org/llmrisk/llm01-prompt-injection/](https://genai.owasp.org/llmrisk/llm01-prompt-injection/)
- 19 SUPERWISE, *Sentinel and Pricing*. Every product statement in part 4. Read October 8, 2026. [superwise.ai/sentinel/](https://superwise.ai/sentinel/) · [superwise.ai/pricing/](https://superwise.ai/pricing/)

## Further reading on superwise.ai

- **The question behind this kit.** *What can and cannot be guaranteed about an AI system?* [superwise.ai/learn/what-can-be-guaranteed/](https://superwise.ai/learn/what-can-be-guaranteed/)
- **Records.** *Can you prove afterwards who did what with AI?* [superwise.ai/learn/proving-who-did-what/](https://superwise.ai/learn/proving-who-did-what/)
- **The words in the contract.** The *AI terms that matter* kit, offered at [superwise.ai/learn/glossary/](https://superwise.ai/learn/glossary/)