IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Claude Skills Security Scanning: HIPAA and ZDR Gaps in Pharma

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-11-10 · claude skills security scanning · claude plugin security · claude skill scanner · hipaa · zero data retention · cmek · agent skills supply chain · mcp security · pharma compliance

Original article: https://intuitionlabs.ai/articles/claude-skill-plugin-scanning-hipaa-zdr-pharma



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes: What Launched and What Is Excluded
4. Threat Model: What a Skill Scanner Can and Cannot See
5. Implementation Considerations and Process Changes
6. Data Analysis and Evidence
7. Frequently Asked Questions (FAQs)
8. Implications and Future Directions
9. Conclusion

Executive Summary

Anthropic announced skill and plugin security scanning for **Claude Enterprise** on **August 6, 2026**, and its release notes label the feature a beta (support.claude.com). The scan checks third-party skills and plugins for malicious content when someone uploads or edits them (support.claude.com), and returns **pass**, **warn**, or **fail**. Most scans finish in about one to two minutes, and results are cached (support.claude.com). Since **October 2, 2026** the control is on by default for Enterprise organizations that had not chosen a setting (support.claude.com).

The coverage is narrower than the feature name suggests. Anthropic excludes skills and plugins that existed before scanning was enabled (support.claude.com), skills created with Claude, skills shared through a connected **Model Context Protocol (MCP)** server, and MCP servers. Organizations using customer-managed encryption keys (CMEK), zero data retention (ZDR), or HIPAA configurations are excluded entirely (support.claude.com). The scanner article describes hook scanning only for hooks shipped inside a plugin. Anthropic states that a pass "isn't a guarantee" of safety (support.claude.com), and that scanning will not catch a skill that behaves in unintended ways without being malicious.

The default change interacts with publishing policy. For Enterprise organizations that never chose a publishing setting, "Open" applies where scanning is on, and "Requires review" applies to all others (support.claude.com). Under "Open", an item that passes the scan is listed without waiting for owner approval. In HIPAA-configured organizations skills are available but not covered under the business associate agreement (BAA) (support.claude.com), and Anthropic tells administrators to review each skill and plugin themselves because scanning is unavailable.

This report builds a coverage and control-gap model rather than a setup guide. Using a hypothetical 134-artifact inventory, the scanner touches 46 artifacts (34.3%), completed-review coverage with human sign-off is 26.9%, and 65.7% of the inventory sits in classes the scanner does not address. In academic work, the regex prefilter of one purpose-built skill gateway caught 135 of 150 malicious benchmark files, but also had a 27.1% false-positive rate on that benchmark (arxiv.org), and Anthropic publishes no efficacy benchmark in the pages reviewed. The recommended posture is to enable scanning where it is available as a pre-filter, keep human

review as the approval gate, and prohibit or tightly limit skills in HIPAA, ZDR, and CMEK tenants until compensating controls (provenance, allowlists, pinned versions, sandbox tests, egress limits, and runtime monitoring) are evidenced.

Introduction and Background

Skills are folders of instructions, scripts, and resources that Claude loads on demand (www.anthropic.com). Because a skill can contain executable code alongside instructions (platform.claude.com), a skill is closer to a software package than to a document. Anthropic's own guidance says to use skills only from trusted sources, meaning ones a user created or obtained from Anthropic (platform.claude.com) (www.anthropic.com), and it warns that a malicious skill can direct Claude to call tools or run code in ways that do not match its stated purpose (platform.claude.com).

For pharma and biotech organizations the stakes are specific. Skills run next to data that may include protected health information (PHI), unpublished research, and records governed by quality systems. The audience for this report is the group that has to decide whether Anthropic's scanner changes the approval process: security architects, Claude Enterprise owners, AI platform teams, privacy officers, research IT, quality partners, and procurement. The question is not whether to scan. It is what the scan covers, what it does not, and what review must fill the gap, especially where the scanner is unavailable.

The report is anchored on three Anthropic Help Center pages accessed on November 10, 2026: the scanner article, the organization provisioning article, and the release notes. Where those pages disagree or are silent, the report says so. For example, the scanner article says an Owner can choose "Allow anyway" for a blocked item, while the organization article says a blocked skill "can't be approved for the organization at this time" (support.claude.com). Anthropic also notes that the override is rolling out gradually, which is one plausible reconciliation, though the pages do not say so.

The analysis relies on third-party and regulator sources, and product behavior is dated to the November 10, 2026 reading because scope and defaults may change quickly.

Key Changes: What Launched and What Is Excluded

What launched on August 6, 2026

The release note for August 6 says Enterprise plans can turn on scanning to check third-party skills and plugins for malicious content at upload or edit. The scanner article specifies the surfaces: **Claude**, **Claude Cowork**, and **Enterprise plugin marketplaces**. Scope includes standalone skills, plugins, and the skills bundled inside a plugin (support.claude.com). There is no extra charge. Admins can narrow the feature to custom roles, and a custom-role capability governs who is subject to scans (support.claude.com).

Content handling is documented in two sentences. Contents are processed in an isolated environment kept separate from normal Claude sessions, and the scanned copy is deleted after the scan, leaving only the result and basic metadata (support.claude.com). The pages reviewed do not say why CMEK, ZDR, and HIPAA organizations are excluded.

The October 2 default change

The scanner article states that scanning is on by default for Enterprise organizations that had not chosen a setting before October 2, 2026. The organization article repeats that the default applies "where available" and that organizations which already chose keep their choice. Two consequences follow for an Enterprise owner who never opened the policy tab. Scanning may now be on without anyone having decided it, and the publishing default has moved with it.

Publishing has three settings. "Requires review" means an owner approves each submission, "Open" publishes without review, and "Off" hides the publish button. For Enterprise plans that had not chosen, Publishing defaults to "Open" where scanning is on and "Requires review" for all others (support.claude.com). Under "Open", items that users publish go straight to the organization library without owner review, and the organization article marks this as important (support.claude.com). Where scanning is on, a new item is not listed until it passes the scan, which usually takes a few minutes. The practical reading is that a scanner pass can substitute for a human approval in a default configuration.

Scan semantics: pass, warn, fail

Table 1 below summarizes the three results as documented, and adds a proposed response for a regulated organization. The response column is analyst recommendation, not Anthropic guidance.

Result	Documented behavior	Who can override	Proposed response (analyst recommendation)
Pass	The item installs normally with no extra message (support.claude.com). A pass means the scan did not find one kind of threat, and Anthropic says it is not a guarantee of safety in every respect.	Not applicable	Treat as an input to review, not as approval. Record the scan result in the artifact manifest and continue to provenance, version pinning, and intended-use checks.
Warn	Claude could not fully verify the item (support.claude.com). It stays usable behind a caution banner that users acknowledge.	Users acknowledge the banner. Owners, Primary Owners, and roles with Libraries set to "Can manage" can choose "Allow anyway" where shown.	Block for regulated data classes until a named reviewer completes static and manual review. A warn is an inability to verify, so the burden of proof moves to the reviewer.
Fail	The scan detected malicious content, and the item is blocked (support.claude.com). A failed submission cannot be approved until the author fixes it and resubmits (support.claude.com).	The uploader cannot override. Owners and "Can manage" roles may allow a version, and an unchanged re-upload returns the same result.	Treat as a security event. Do not use "Allow anyway" without a documented exception (see the exception process below). Quarantine the source and preserve the package for analysis.

Two details matter operationally. First, an override applies only to the version it was granted on, and each new version is scanned again (support.claude.com). Second, results are cached, so re-uploading the same content returns the earlier result almost instantly. Together these mean scan outcomes are tied to content, not to a publisher's reputation, which is a useful property for a fixed-version approval model.

The coverage-gap matrix

Table 2 crosswalks artifact classes against the scan coverage documented in the scanner article. "Not described" means the Anthropic pages reviewed do not address the class, which is different from "excluded". These entries are the analyst's reading of silence in those pages, not documented exclusions.

Artifact class	Upload or delivery path	Scan coverage (per Anthropic)	Review that must fill the gap
Third-party skill uploaded after enablement	Member or owner upload, or edit	Scanned	Provenance check, pinned version, intended-use gate
Third-party plugin, including bundled skills	Upload to org settings or plugin marketplace	Scanned, including bundled skills	Same as above, plus review of any MCP server or hook it declares
Hooks shipped inside a plugin	Plugin package	Scanned	Manual review of every command hook, which run with full user permissions (code.claude.com)
Hooks outside a plugin	Managed settings, repository configuration	Not described in the scanner article	Settings review, and restrict non-plugin customization where possible (code.claude.com)
Items present before scanning was enabled	Existing library or user uploads	Excluded	Back-scan and re-review as a one-time inventory exercise
Skills created with Claude	Created in a Claude session	Excluded, relies on built-in safeguards	Control through the User-created skills toggle (support.claude.com)
Skills delivered through MCP	Connected MCP server	Excluded, follows the MCP trust model	Treat as part of the MCP server review
MCP servers	Connector or local server	Not scanned at this time	Server allowlist, scoped tokens, and sandboxing; Anthropic says it does not security-audit or manage any MCP server (code.claude.com)
Any item in a CMEK, ZDR, or HIPAA tenant	Any	Scanning not available	Full manual review under the tenant's change control
Repository-synced marketplaces	GitHub sync into org marketplace	Not described; a sync replaces all plugins with the repo state (support.claude.com)	Review the repository, not only the listing

The matrix has an asymmetry worth stating plainly. The classes the scanner covers are the ones with the clearest human chokepoint, an upload or an edit. The classes it excludes are those that arrive without such a chokepoint: items already installed, content generated inside a session, and anything delivered by a server at runtime.

Publishing policy and fixed-version review

The organization article gives reviewers real tools. A reviewer sees who submitted the item, its scan result, every file, and a diff against the published version for updates (support.claude.com). The reviewed version is fixed: author edits do not reach the organization until a new version is submitted and approved. When an owner approves a new version, everyone using the item gets the update automatically (support.claude.com). Owners' own submissions skip review.

Update drift is the second half of the version story. A shared skill updates for recipients at next use when the owner saves a new version. Inside Claude Code, plugin auto-update is off by default for third-party marketplaces and on for some official ones (code.claude.com), and when it is on, the files a user reviewed can change on disk (code.claude.com). Group scoping is possible by bundling skills into a plugin assigned to a group, but custom-role permissions are additive, so one role cannot remove what another grants (support.claude.com). Where a member belongs to several groups with different plugin settings, the most permissive setting applies (support.claude.com).

Threat Model: What a Skill Scanner Can and Cannot See

A usable threat model for skills has five classes, each with a different detection profile.

- **Malicious instructions.** Natural-language directions that redirect the agent. **OWASP** notes that it is unclear whether any prevention of prompt injection is fool-proof (genai.owasp.org).
- **Package and code execution.** Bundled scripts and plugin commands. A **Claude Code** plugin can execute arbitrary code with the user's privileges (code.claude.com), and command hooks are shell commands that run with the user's full permissions (code.claude.com).
- **Exfiltration.** Moving data out. On claude.ai and in Claude Code, skill network access can match any other program on the user's computer (platform.claude.com), while skills on the API run in a container without network access (platform.claude.com).
- **Privilege misuse.** Using granted access beyond purpose. OWASP's Excessive Agency entry covers extensions "sometimes referred to as tools, skills or plugins" (genai.owasp.org) and applies regardless of what causes the model to malfunction (genai.owasp.org).
- **Unsafe but non-malicious behavior.** Anthropic is explicit that a pass does not mean the item will not behave in unintended ways. This class is outside what the scanner is described as checking.

A 2026 taxonomy of agent-skill threats organizes seven categories and nineteen scenarios across three attack layers (arxiv.org), and attributes the most severe to structural properties such as the absence of a data-instruction boundary and a single-approval persistent trust model (arxiv.org). The same logic explains why supply-chain controls matter even when a scan passes: dependency change after approval is a documented risk, since even trustworthy skills can be compromised if external dependencies change (platform.claude.com).

What the research says about scanner efficacy

Anthropic does not publish detection-rate benchmarks in the pages reviewed, and this report found no published independent evaluation of Anthropic's scanner. Research on other skill scanners is informative about the category, with the caveat that results do not transfer automatically.

- **SkillGate (July 2026)** is a screening gateway built as a hybrid regex prefilter plus a large-language-model judge (arxiv.org), with 530 regex patterns (arxiv.org). Its prefilter alone caught 135 of 150 malicious benchmark files but also flagged 406 benign ones, a 27.1% false-positive rate the authors call far above deployable constraints (arxiv.org), and the hybrid design cut LLM input tokens by 77% (arxiv.org). The authors state that the 150 malicious files are hand-crafted exemplars, not in-the-wild samples (arxiv.org), that the judge is nondeterministic (arxiv.org), and that adaptive adversaries are out of scope (arxiv.org).
- **A large-scale static study** reported that **26.1%** of analyzed skills contain at least one vulnerability (arxiv.org), with a detector at 86.7% precision and 82.5% recall on a validated sample (arxiv.org). Its error analysis calls evasion patterns fundamental limits of static and single-pass analysis (arxiv.org).
- **Adversarial evasion work** reports that a multimodal attack achieved 78 to 100 percent attack success against every one of six baseline scanners (arxiv.org), and that splitting intent across several plausible skills reached an average 96.0% success (arxiv.org).

Read together, these studies support a modest conclusion: scanners reduce exposure to unsophisticated or known patterns, and the literature's own authors position them as one layer. None of this evaluates Anthropic's service, whose methods Anthropic does not describe.

Implementation Considerations and Process Changes

Compensating controls

Each control below is verifiable against a vendor or standards source.

- **Provenance. NIST**'s secure software development framework asks organizations to collect and maintain provenance data for all components of each release (nvlpubs.nist.gov). **SLSA** defines provenance as what entity built the artifact, what process they used, and what the inputs were (slsa.dev), and notes that its lowest level is trivial to bypass or forge (slsa.dev). OWASP's agentic list adds a mitigation to verify provenance before install or activation and reject unsigned or unverified components (genai.owasp.org).
- **Allowlists.** Claude Code's managed settings decide which marketplace sources plugins may come from (code.claude.com), applied before anything downloads and again at session start (code.claude.com). The allowlist does not restrict entries inside an allowed marketplace, and plugin keys apply to every user who receives the settings (code.claude.com).
- **Pinned versions and checksums.** A plugin source can be pinned to a full 40-character commit SHA (code.claude.com), and Claude Code refuses an archive install whose digest does not match (code.claude.com). GitHub's guidance notes a tag can be moved or deleted (docs.github.com), and that pinning to a full SHA requires an attacker to produce a SHA-1 collision (docs.github.com).
- **Static and manual review. Cisco**'s scanner detects prompt injection, data exfiltration, and malicious code patterns (github.com) using static, behavioral, and language-model analysis (github.com), and its documentation advises pairing results with manual code review for high-risk deployments (github.com).

HOL's plugin scanner checks security posture such as secrets, action pinning, and lockfiles (hol.org). **Snyk**'s agent-scan sends component information to an analysis API (github.com), which matters if skill content is confidential.

- **Secrets and dependencies. Gitleaks** finds secrets in repositories and files (github.com), TruffleHog can confirm whether a found secret is live (github.com), and OSV-Scanner finds known vulnerabilities in dependencies (google.github.io).
- **Sandbox tests. gVisor** places a per-sandbox application kernel between the workload and the host (gvisor.dev), and a Docker container with no network driver isolates its networking stack (docs.docker.com).
- **Least privilege.** OWASP advises limiting the extensions an agent may call to the minimum necessary (genai.owasp.org) and requiring human approval for high-impact actions (genai.owasp.org). In Claude Code, permission rules are evaluated deny first (code.claude.com), and no command-line argument can override a managed rule (code.claude.com). The managed MCP allowlist is a separate setting that other permission flags do not enforce (code.claude.com).
- **Egress controls.** The Claude Code sandbox sends network traffic through a proxy that checks each host against allowed domains, which start empty (code.claude.com). Built-in tools such as WebFetch follow permission rules rather than the sandbox (code.claude.com), so both layers need configuration.
- **Secret isolation.** The sandbox's default read access covers most of the machine, including credential files (code.claude.com), so credentials should not live on hosts where skills run.
- **Runtime monitoring and revocation.** Anthropic's **November 5, 2026** security update describes inference hooks that send prompts and tool calls to a customer's data security tools (claude.com). Skill-sharing events appear in the audit log, but not skill contents (support.claude.com). OWASP recommends an emergency revocation mechanism that can disable specific tools quickly (genai.owasp.org).

A fixed-version provenance manifest ties these together. For every approved artifact, record:

- **Identity.** Name, type (skill, plugin, hook, MCP server), and publisher of record.
- **Source reference.** Repository URL, full commit SHA, or archive SHA-256 digest.
- **Content inventory.** File list with hashes, declared dependencies, and any network endpoints, in a format such as **CycloneDX**, which describes first-party and third-party components (cyclonedx.org).
- **Signatures. Sigstore**-style signature and transparency-log entry where available (docs.sigstore.dev).
- **Scan result.** Anthropic scan outcome and date, plus any open-source scanner outputs.
- **Human review.** Reviewer, date, intended use, data classes permitted, and tools granted.
- **Approval scope.** Groups, default access (available, installed by default, or required), and expiry date.
- **Rollback.** The prior approved version and the revocation owner.

The CISA-led 2026 minimum elements for a software bill of materials add component hash, license, tool name, and generation context (content.govdelivery.com), a convenient checklist for the content inventory.

HIPAA, ZDR, and CMEK control crosswalk

Table 3 crosswalks the three excluded configurations against what Anthropic documents about skills, and the minimum compensating controls proposed here.

Configuration	What Anthropic documents	Skill and plugin status	Minimum compensating controls (analyst proposal)
HIPAA, Enterprise chat	Only a **Primary Owner** can accept the BAA and enable HIPAA (support.claude.com), and the change cannot be reversed from organization settings.	Skills are available but not covered under the BAA (support.claude.com). Skills and sharing are off by default. Scanning unavailable.	Keep skills off for PHI workflows, or allow only owner-provisioned, human-reviewed, pinned skills. Set Publishing to "Requires review".
HIPAA configuration for Claude Code and Cowork (local mode)	Enabling HIPAA alone does not bring Claude Code or Cowork under the BAA (support.claude.com).	Applying it turns off Cowork, the Code section of Desktop, skills, and Claude in Chrome. Plugins and skills stay outside the BAA even when turned on (support.claude.com).	Treat any re-enabled skill as outside BAA scope. Prohibit PHI in sessions that use skills. Re-enable only for approved, reviewed items.
ZDR	On the API, ZDR means Anthropic does not store prompts or responses at rest (platform.claude.com). Claude Code ZDR cannot be enabled from admin settings (code.claude.com).	The API skills feature is not ZDR-eligible and its data follows standard retention (platform.claude.com). The Agent Skills overview states it is not covered by ZDR arrangements. ZDR does not extend to third-party tools or MCP servers (code.claude.com).	Do not use skills in ZDR-dependent workflows. If used elsewhere, keep them in a separate organization from ZDR workloads.
CMEK	CMEK on Enterprise encrypts chat content, including skills and plugins (platform.claude.com).	Scanning cannot be turned on, and skills and plugins install without a scan (platform.claude.com).	Full manual review and allowlisting. Encryption protects stored content, not the behavior of an installed skill.
CMEK plus ZDR	On Claude Enterprise an organization can use both (platform.claude.com).	Same exclusions apply.	Combine the controls above.
Standard Enterprise (no excluded configuration)	Standard plans have no BAA coverage until a Primary Owner acts.	Scanning on by default since October 2 for organizations that had not chosen.	Scanner as pre-filter. Publishing "Requires review" for regulated roles.

Two cautions on *Table 3*. First, the controls column is analytical, not a compliance determination, and any HIPAA conclusion needs the organization's privacy officer and counsel. Second, the HHS Security Rule is technology neutral and does not dictate specific measures (www.hhs.gov), but a risk analysis affects every safeguard (www.hhs.gov). A HIPAA tenant's risk analysis should therefore name skills and plugins as assets. HHS's proposed Security Rule update would also require a technology asset inventory and a network map of ePHI movement (www.hhs.gov), though HHS states the current rule remains in effect while the rulemaking is pending (www.hhs.gov).

Pharma data-class and intended-use gates

A skill's permitted use should depend on both the data it touches and the decision it informs. A five-class scheme works for most pharma programs:

- **Class 0, public.** Any approved skill.
- **Class 1, internal non-GxP.** Scanner pass plus provenance check.
- **Class 2, confidential R&D or commercial.** Human review, pinned version, egress allowlist.
- **Class 3, PHI or personal data.** Only in a tenant under a BAA, with skills excluded from BAA scope treated as prohibited unless the privacy officer accepts a documented risk. HHS notes that de-identified data under either method still retains some risk of identification (www.hhs.gov), and Safe Harbor identifiers include device identifiers (www.hhs.gov) and any other unique identifying code (www.hhs.gov), which can appear in tool logs.
- **Class 4, GxP records or regulated decisions.** Full intended-use assessment, quality approval, and audit trail.

For Class 4, the established frameworks are risk based and tied to intended use. FDA's computer software assurance guidance, reissued on **February 3, 2026**, defines assurance as a risk-based approach (www.fda.gov), though its scope is device production and quality-system software, so for drug programs it is an analogy rather than a direct requirement (www.fda.gov). **GAMP 5** second edition says the regulated company determines the type and level of assurance by fitness for intended use (ispe.org). **ICH Q9(R1)** states that effort, formality, and documentation should be commensurate with risk (health.ec.europa.eu).

Records rules apply as well. **21 CFR Part 11** requires validation of systems (www.ecfr.gov), and secure, time-stamped audit trails (www.ecfr.gov). The European Commission's draft **Annex 11** requires quality risk management throughout a computerised system's lifecycle (health.ec.europa.eu) and sets validation effort by intended use and risk (health.ec.europa.eu). Draft **Annex 22** states that it does not apply to generative AI and large language models (health.ec.europa.eu), and that for non-critical GMP applications qualified personnel remain responsible for output suitability (health.ec.europa.eu). The consultation closed on **October 7, 2025** (health.ec.europa.eu), and this report found no final adoption notice as of the reading date. FDA's draft guidance on AI for regulatory decision-making scales credibility activities to model risk and context of use (www.fda.gov) and excludes AI used only for operational efficiencies such as drafting (www.fda.gov).

The implication: an agent skill that drafts internal text may sit outside these frameworks, while a skill that touches a record or a quality decision must be assessed for intended use and change-controlled. Draft Annex 22 requires change control over the model, the system, and the process (health.ec.europa.eu), which maps onto version pinning and update review for skills.

Pilot and negative-test pack

A pilot should answer two questions: what does the scanner catch in this tenant, and what do the other controls catch when it does not? Use synthetic data, canary secrets, and a non-production organization where possible. Confirm with the Anthropic account team and legal that test samples are permitted before uploading deliberately unsafe test content.

- **Prompt injection.** A skill whose instructions tell the agent to ignore prior directions and disclose its system context. Expected: scan warn or fail, or runtime refusal. OWASP recommends human-in-the-loop controls for privileged operations (genai.owasp.org).
- **Data exfiltration.** A script that posts a canary file to an unapproved host. Expected: egress proxy block and an alert, since sandbox traffic is checked against allowed domains (code.claude.com).
- **Secret access.** A skill that reads a planted fake credential file. Expected: no access, because default read access reaches credential files (code.claude.com) unless denied.
- **Dependency install.** A skill that installs an unpinned package at run time. Expected: block, because dependencies can change after review.
- **Obfuscation.** Encoded payloads, hidden characters, and instructions placed inside images. Research on evasion shows encoding-based false negatives remain.
- **Unsafe tool scope.** A skill that requests broad tool permissions. Expected: denial under managed deny rules (code.claude.com).
- **Update drift.** Approve version one, then push version two to the source. Expected: users remain on the approved version and the new one enters review.
- **MCP delivery.** A skill delivered through a test MCP server. Expected: not scanned, so detection must come from server allowlists and review. MCP guidance says clients should warn that servers run with the same privileges as the client (modelcontextprotocol.io).

Record pass criteria per test and rerun the pack at each policy change and quarterly. A missed detection is a finding about the layer, not a failure of the pilot.

Exception and incident process

An exception is needed whenever a warned or failed version is allowed, or an excluded class is used with regulated data. The record should name the requester, the version or digest, the data classes, the reviewer, the compensating controls, and an expiry. Anthropic limits "Allow anyway" to Owners, Primary Owners, and roles with Libraries set to "Can manage", and allowing a version removes the banner for everyone with access to it (support.claude.com), so the decision is organization-wide in effect. An allowance can be undone with "Revoke", and an owner can remove an item from the library.

If a published item is suspected of misbehaving, the response is: remove or disable the item, hold dependent workflows, preserve the package and logs, review the audit log for sharing events, and notify the privacy officer if PHI could have been involved. NIST's generative AI profile recommends incident response plans for third-party generative AI technologies (nvlpubs.nist.gov). Plugins set to "Required" cannot be turned off by users (support.claude.com), so owners must be able to remove them centrally.

Enable, limit, or prohibit decision tree

1. **Is the tenant under HIPAA, ZDR, or CMEK?** If yes, scanning is unavailable. Default to **prohibit** for user-created and third-party skills. Allow **limited** use of owner-provisioned, fully reviewed, pinned skills for Classes 0 to 2 only. Keep Publishing at "Requires review".
2. **If no, will the artifact class be scanned?** Existing items, Claude-created skills, MCP-delivered skills, and MCP servers are not. For these, **limit** to reviewed and allowlisted items.

3. **Does the skill touch Class 3 or 4 data?** If yes, require human review and a fixed version regardless of scan result. If the BAA does not cover the surface, **prohibit**.
4. **Is Publishing set to "Open"?** For regulated roles, switch to "Requires review" so a scan pass cannot be the only gate.
5. **Otherwise, enable** scanning as a pre-filter with the manifest, and rerun the negative-test pack at each change.

Data Analysis and Evidence

Coverage calculation (Hypothetical Example)

The prescribed metric uses the reader's own inventory: reviewed applicable artifacts divided by total inventoried artifacts, with excluded classes reported separately and never counted as passes. The following worked example is a **hypothetical** inventory of 134 artifacts in a standard Enterprise tenant that has no CMEK, ZDR, or HIPAA configuration.

- **Pre-enablement third-party items:** 38 (excluded from scanning).
- **Post-enablement third-party uploads and edits:** 46 (scanned: 39 pass, 5 warn, 2 fail).
- **Skills created with Claude:** 21 (excluded).
- **MCP-delivered skills:** 9 (excluded).
- **MCP servers:** 12 (excluded).
- **Hooks outside plugins:** 8 (not described).

Scanner-touched artifacts are 46 of 134, or **34.3%**. A completed review requires the scan result plus a documented human review. Suppose human review is complete for 31 of the 39 passes and all 5 warns, while the 2 fails remain blocked and out of the numerator. Reviewed coverage is (31 + 5) / 134 = **26.9%**. The excluded classes total 38 + 21 + 9 + 12 + 8 = 88 artifacts, or **65.7%** of inventory, and are reported separately. If compensating review is completed on 24 of those 88, combined reviewed coverage is (36 + 24) / 134 = **44.8%**. In a HIPAA, ZDR, or CMEK tenant the scanner term is zero by construction, and the full inventory depends on compensating review. The numbers are illustrative; readers should compute the metric from their own inventory.

Ecosystem and risk statistics

The following figures are aggregate and come from originators. They indicate exposure context, not the behavior of any specific product.

- **Skill ecosystem measurement.** A study collected **42,447** skills from two public marketplaces (arxiv.org). In its analysis, data exfiltration (13.3%) and privilege escalation (11.8%) were the most prevalent patterns (arxiv.org), and 5.2% of skills showed high-severity patterns suggesting malicious intent (arxiv.org).
- **Behaviorally confirmed malicious skills.** A separate analysis of **98,380** skills (arxiv.org) identified **157** with confirmed malicious behavior, about 0.16%, and its authors note a 60-second dynamic window may miss delayed payloads (arxiv.org).

- **MCP tool poisoning.** A benchmark built on 45 live MCP servers and 353 tools (arxiv.org) concluded that safety alignment is ineffective against malicious actions that use legitimate tools (arxiv.org). A cross-framework measurement found metadata scanning alone is a partial control (arxiv.org).
- **Breach cost context.** IBM's 2026 **Cost of a Data Breach Report** puts the global average at **USD 4.99 million** (www.ibm.com), based on 602 organizations (newsroom.ibm.com), and healthcare at **USD 6.64 million**, the highest of any industry for the 15th consecutive year (www.ibm.com).
- **AI access controls.** In IBM's 2026 data, 92% of organizations reporting an AI-related breach lacked proper AI access controls (www.ibm.com), and only 40% of organizations reported using access controls on AI models and data (www.ibm.com). IBM lists compromised APIs, applications, or plug-ins as a cause in 27% of breaches targeting AI (newsroom.ibm.com).
- **Third-party involvement. Verizon**'s 2026 Data Breach Investigations Report puts third-party involvement at **48%** of breaches (www.verizon.com), for incidents between November 1, 2024, and October 31, 2025 (www.verizon.com).
- **Healthcare reporting.** HHS's report to Congress for calendar year 2024 records **663** breach notifications affecting 500 or more individuals (www.hhs.gov), covering approximately 242,908,056 individuals (www.hhs.gov).
- **Package ecosystem. Sonatype** identified more than **454,600** new malicious open-source packages in 2025 (www.sonatype.com).
- **Analyst view. Gartner** names AI application compromise, prompt injection, and software supply chains among four critical threats (www.gartner.com).

These figures do not isolate Claude skills and should not be read as scanner efficacy evidence. The 0.16% and 5.2% figures are not comparable, because the studies define "malicious" differently and sample different registries.

Frequently Asked Questions (FAQs)

Is Claude HIPAA compliant?

Anthropic offers a HIPAA-ready Enterprise configuration that is limited to Enterprise plans and requires the Primary Owner to accept the BAA. Features are sorted into covered, available but not covered, and disabled, and skills sit outside BAA coverage. Compliance remains a determination for the covered entity or business associate, so consult counsel and the privacy officer.

Does Claude zero data retention work with HIPAA?

They are different mechanisms. Anthropic describes HIPAA-ready API access as an alternative to ZDR (platform.claude.com), and its Claude Code HIPAA configuration is meant to replace ZDR for local mode (support.claude.com). Even with ZDR or HIPAA, flagged content can be retained for up to two years (platform.claude.com).

What do Claude plugin security scanning and malicious Claude skills detection cover?

Third-party skills, plugins, skills bundled in plugins, and hooks shipped in a plugin, at upload or edit. Existing items, Claude-created skills, MCP-delivered skills, and MCP servers are outside it. Anthropic does not describe methods or accuracy in the pages reviewed, while one academic study of public skills (arXiv 2601.10338) reports a detection methodology with 82.5% recall at 86.7% precision on a validated sample (arxiv.org); that result does not measure Anthropic's scanner.

What are the main Claude skills security risks, and how is Claude Code plugin marketplace security handled?

Skills can carry executable code, plugins run with the user's privileges (code.claude.com), and updates can change reviewed files (code.claude.com). A marketplace name shows who publishes the catalog, not what each plugin does, and administrators can restrict marketplace sources with a managed allowlist (code.claude.com).

What about Claude skills and pharma compliance, and agent skills supply chain security?

Treat each skill by intended use and data class. ICH Q9(R1) and GAMP 5 both scale effort to risk (health.ec.europa.eu) (ispe.org). For supply chain security, use provenance records (nvlpubs.nist.gov), allowlists, pinned versions (code.claude.com), a component inventory (genai.owasp.org), and a revocation plan (genai.owasp.org).

Implications and Future Directions

Several uncertainties will shape how long this analysis stays current. First, Anthropic labels the scanner a beta, says MCP servers are not scanned "at this time", and says the override is rolling out gradually, so the exclusion list may change. Re-read the scanner article at each procurement review instead of carrying forward this report's matrix.

Second, regulators are moving toward explicit AI guidance for GxP. The European Commission's draft Annex 22 emphasizes the intended use of the model, performance metrics, training data quality, and test data management (health.ec.europa.eu), and it limits critical applications to static models (health.ec.europa.eu).

Third, agent security is shifting from perimeter checks toward runtime enforcement. OWASP's agentic guidance observes that agentic ecosystems compose capabilities at runtime, which expands the attack surface beyond static dependencies (genai.owasp.org), and Anthropic's inference hooks point the same way (claude.com).

HHS has also proposed, but not finalized, a Security Rule update that would require a technology asset inventory, so teams that inventory skills and plugins now are better placed for any final text (www.hhs.gov). The scanner is best framed as a useful pre-filter and not an approval. The practical agenda is to inventory artifacts, measure reviewed coverage with excluded classes reported separately, align publishing policy with data classes, and wire compensating controls into the quality system. Procurement teams should ask Anthropic for documented scope commitments, detection methodology, and a change-notification path for scanner coverage.

Conclusion

Anthropic's skill and plugin scanning is a real but bounded control. It reviews third-party skills and plugins at upload or edit, returns pass, warn, or fail, and is on by default for Enterprise organizations that had not chosen a setting as of October 2, 2026. It does not cover items already present, Claude-created skills, MCP-delivered skills, MCP servers, or any organization using CMEK, ZDR, or HIPAA configurations, and a pass is explicitly not a guarantee of safety.

For pharma and biotech, the decision rests on three moves. Keep human approval as the gate, particularly because default publishing can let a scan pass stand in for review. Report coverage as reviewed artifacts over total inventory with excluded classes separate. And treat HIPAA, ZDR, and CMEK tenants as unscanned by design, applying provenance, allowlists, pinned versions, sandbox tests, least privilege, egress controls, secret isolation, and runtime monitoring. Because the pages reviewed publish no detection benchmarks, the scanner's marginal protection cannot be quantified, which is a further reason to keep compensating review in place. Facts here reflect Anthropic documentation read on November 10, 2026 and should be reverified before each decision. Teams that want help translating this model into a quality-system procedure can contact IntuitionLabs.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://support.claude.com/en/articles/12138966-release-notes
2. https://support.claude.com/en/articles/15927065-get-started-with-skill-and-plugin-scanning
3. https://support.claude.com/en/articles/13119606-provision-and-manage-skills-for-your-organization
4. https://support.claude.com/en/articles/8114513-business-associate-agreements-baa-for-commercial-customers
5. https://arxiv.org/html/2607.25619v1
6. https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-skills
7. https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview
8. https://support.claude.com/en/articles/13930452
9. https://code.claude.com/docs/en/hooks
10. https://code.claude.com/docs/en/plugins/org
11. https://code.claude.com/docs/en/security
12. https://support.claude.com/en/articles/13837433
13. https://code.claude.com/docs/en/discover-plugins
14. https://code.claude.com/docs/en/plugins/security
15. https://genai.owasp.org/llmrisk/llm01-prompt-injection
16. https://genai.owasp.org/llmrisk/llm062025-excessive-agency
17. https://arxiv.org/html/2604.02837
18. https://arxiv.org/html/2601.10338
19. https://arxiv.org/html/2606.18198

20. https://arxiv.org/html/2608.09732
21. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218.pdf
22. https://slsa.dev/spec/v1.0/levels
23. https://genai.owasp.org/download/52117?tmstv=1765059207
24. https://code.claude.com/docs/en/plugins/marketplace-reference
25. https://docs.github.com/en/actions/reference/security/secure-use
26. https://github.com/cisco-ai-defense/skill-scanner
27. https://hol.org/docs/libraries/ai-plugin-scanner/overview
28. https://github.com/snyk/agent-scan
29. https://github.com/gitleaks/gitleaks
30. https://github.com/trufflesecurity/trufflehog
31. https://google.github.io/osv-scanner
32. https://gvisor.dev/docs
33. https://docs.docker.com/engine/network/drivers/none
34. https://code.claude.com/docs/en/permissions
35. https://code.claude.com/docs/en/managed-mcp
36. https://code.claude.com/docs/en/sandboxing
37. https://claude.com/resources/articles/claude-enterprise-security-updates
38. https://cyclonedx.org/specification/overview
39. https://docs.sigstore.dev/cosign/signing/overview
40. https://content.govdelivery.com/accounts/USDHSCISA/bulletins/422a7eb
41. https://support.claude.com/en/articles/13296973-hipaa-ready-enterprise-plans
42. https://support.claude.com/en/articles/17318731
43. https://platform.claude.com/docs/en/manage-claude/api-and-data-retention
44. https://code.claude.com/docs/en/zero-data-retention
45. https://platform.claude.com/docs/manage-claude/cmek-coverage-claude-enterprise
46. https://platform.claude.com/docs/en/manage-claude/cmek-data-retention
47. https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html
48. https://www.hhs.gov/hipaa/for-professionals/security/hipaa-security-rule-nprm/factsheet
49. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
50. https://www.fda.gov/media/188844/download
51. https://ispe.org/pharmaceutical-engineering/january-february-2023/what-you-need-know-about-gampr-5-guide-2nd-edition
52. https://health.ec.europa.eu/system/files/2023-10/mp_ich_q9guideline-ra_en.pdf
53. https://www.ecfr.gov/api/renderer/v1/content/enhanced/current/title-21?part=11
54. https://health.ec.europa.eu/document/download/40231f18-e564-4043-94de-c031f813d38b_en?filename=mp_vol4_chap4_annex11_consultation_guideline_en.pdf

55. https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en?filename=mp_vol4_chap4_annex22_consultation_guideline_en.pdf
56. https://health.ec.europa.eu/consultations/stakeholders-consultation-eudralex-volume-4-good-manufacturing-practice-guidelines-chapter-4-annex_en
57. https://www.fda.gov/media/184830/download
58. https://modelcontextprotocol.io/docs/tutorials/security/security_best_practices
59. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
60. https://arxiv.org/html/2602.06547
61. https://arxiv.org/html/2508.14925
62. https://arxiv.org/html/2611.03555
63. https://www.ibm.com/think/x-force/2026-cost-of-a-data-breach-ai-adversaries-enterprise-risk
64. https://newsroom.ibm.com/2026-07-29-ibm-study-one-in-four-malicious-breaches-are-ai-enabled,-costing-companies-6-million-on-average
65. https://www.ibm.com/think/topics/data-breach
66. https://www.verizon.com/about/news/breach-industry-wide-dbir-finds
67. https://www.verizon.com/business/resources/reports/dbir
68. https://www.hhs.gov/sites/default/files/breach-report-to-congress-2024.pdf
69. https://www.sonatype.com/state-of-the-software-supply-chain/2026/open-source-malware
70. https://www.gartner.com/en/newsroom/press-releases/2026-06-02-gartner-identifies-four-critical-threats-requiring-urgent-improvements-from-cybersecurity-leaders
71. https://genai.owasp.org/llmrisk/llm032025-supply-chain

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.