CLAUDE MANAGED AGENTS PRIMER

Managed Agents Security

Network, credentials, approvals, and budgets

A visual primer on the documented settings that limit what a Claude Managed Agents deployment can reach and do, and how to verify them. Independent analysis; not sponsored or endorsed by Anthropic.

FIRST-PARTY DOCUMENTATION / INDEPENDENT LUMINITY SYNTHESIS



The guide's settings limit what a Managed Agent can reach and do

Anthropic's security guide lists them most important first. This primer covers eight areas, grouped by what each limits (Luminity's grouping).

AT A GLANCE

8
areas of the guide this primer covers (Luminity's count)

3
approval policies: always_allow, auto, always_ask

3
vault credential types: mcp_oauth, static_bearer, environment_variable

6
steps to stop a session that should not be running

What the guide has you set

DOCS

Run on Anthropic's cloud unless tools must run on your infrastructure, limit the network to the hosts the agent needs, keep secrets in vaults, name its tools, and cap spend.

What the settings do not do

DOCS

They lower the risk and do not remove it. They do not cover custom tools, MCP servers, or self-hosted infrastructure. Limit each key at the service that issued it.

BEFORE YOU RELY ON IT

- ! Beta (managed-agents-2026-04-01). Docs verified Nov. 27, 2026 — STATUS
- ! Not HIPAA eligible. Not Legacy ZDR eligible, except self-hosted — ELIGIBILITY
- ! The guide also covers IP allowlisting, workload identity, Google Cloud IAM — NOT COVERED

IN THIS PRIMER

Key Points

01 **Reach**
Network limits, the placeholder key in the sandbox, and vault credentials.

02 **Act**
Approval policies, paused calls, and untrusted input.

03 **Cap and stop**
Session budgets, what happens at the cap, and the stop sequence.

04 **Watch and keep**
Webhooks, the offline reviewer, retention, and the self-hosted boundary.

A decision path closes the primer. The groups and the path are Luminity's; every setting is from the docs. Informational, not compliance advice.

© 2026 Luminity Digital, Inc. | Independent analysis of Anthropic public documentation; not sponsored or endorsed by Anthropic.

Source: Claude Platform Docs: secure-deployment, secure-deployment-risks, api-and-data-retention · platform.claude.com/docs/en/ · beta · verified Nov. 27, 2026

Anthropic and Claude are trademarks of Anthropic PBC.

Limit what an agent can reach, and keep keys out of its sandbox

The guide's settings for where it runs, what it can contact, and how secrets are delivered. Statements are the docs'; the order is Luminity's.

01

What does the guide have you set?

- ✓ Run on cloud unless tools need your infra — WHERE
- ✓ Network: limited, only the hosts it needs — NETWORK
- ✓ Web tools: allowed_domains on each tool — WEB TOOLS
- ✓ Secrets: in a vault, out of direct reach — VAULT
- ! An org admin's web settings apply too — ORG

Controls are the docs'; their order is Luminity's

02

How does each credential type reach the agent?

Docs: Authenticate with vaults | **Recommended: headers only**

mcp_oauth	static_bearer	environment_variable	Injection
Keyed by MCP server URL	A fixed token, one server	Keyed by secret_name	environment_variable only
REFRESHED	NO REFRESH	PLACEHOLDER	LOCATION
Injected when the agent connects to that server. Add a refresh block and Anthropic refreshes the token	**Injected on connect, such as an API key for one MCP server URL**	**The sandbox holds a placeholder; the real value is swapped in at egress, only on allowed_hosts**	**Omit it on create and header and body are both on. Console-created credentials are header-only**

03

Where the docs say it stops

Unlisted host — VAULTS

No key is sent. The request is blocked only if the environment also disallows the host. A credential with unrestricted networking is swapped in on any host the environment reaches.

Workspace scope — VAULTS

Vaults are workspace-scoped: any API key with workspace access can reference them. Keep workloads of different trust in separate workspaces.

Org web settings — WEB TOOLS

An admin's Console settings also apply to web_search and web_fetch. A call must pass every list, so the org's and the agent's lists intersect.

Self-hosted — VAULTS

environment_variable credentials are not yet supported with self-hosted sandboxes.

READING RULE · Settings and limits: the docs · Order and grouping: Luminity's · Docs verified Nov. 27, 2026 · Informational, not compliance advice

© 2026 Luminity Digital, Inc. | Independent analysis of Anthropic public documentation; not sponsored or endorsed by Anthropic.

Source: Claude Platform Docs: secure-deployment, vaults, tools-web-restrictions · platform.claude.com/docs/en/managed-agents/ · beta · verified Nov. 27, 2026

Anthropic and Claude are trademarks of Anthropic PBC.

Choose who approves each tool, and fence off untrusted input

Policies are set per toolset or per tool. The agent toolset defaults to always_allow and MCP toolsets to always_ask; no toolset uses auto by default.

01

What are the three policies?

- ✓ always_allow: runs with no confirmation — ALLOW
- ! auto: server decides; not a human check — AUTO
- ✓ always_ask: a person approves each call — ASK
- ! Custom tools: your code decides — NOT GOVERNED
- ! Agent toolset default: bash included — DEFAULT

Policy names and defaults are the docs'

02

What does the server do under auto?

Docs: Permission policies | **One outcome per call**

Allow	Deny	Ask	Answer
Runs	Does not run	Waits for you	user.tool_confirmation
JUDGED SAFE	JUDGED HIGH-RISK	NO DETERMINATION	YOUR CLIENT
Runs as under always_allow. evaluated_permission: allow	**The agent gets an error result. Your client cannot override it**	**The session pauses, as under always_ask, and waits indefinitely**	**One confirmation per blocking event; a 400 if it was not ask**

03

Where the docs say to be careful

Your words count — AUTO

Your user.message text counts as intent and can lead the server to allow a call. Tool results, fetched pages, and MCP responses are not treated as your intent.

Relayed text — RULE

If the session relays end-user text, put bash on always_ask. Have a person approve every call to a tool that can move money.

Split the reader — ISOLATION

A fetcher reads the web, a lead routes, a writer holds change tools. They share one sandbox and vault credentials; use two sessions for a stronger split.

A paused session — WAITS

With other work open, a session can stay running and send no session.status_idle. Answer the agent.tool_use event itself.

READING RULE · Policies and outcomes: the docs · Grouping and tiles: Luminity's · Docs verified Nov. 27, 2026 · Informational, not compliance advice

© 2026 Luminity Digital, Inc. | Independent analysis of Anthropic public documentation; not sponsored or endorsed by Anthropic.

Source: Claude Platform Docs: permission-policies, secure-deployment · platform.claude.com/docs/en/managed-agents/ · beta · verified Nov. 27, 2026

Anthropic and Claude are trademarks of Anthropic PBC.

Cap spend, and know how to stop a session

A budget pauses the session at the cap. In-flight requests can finish, so spend can pass it. The stop sequence is the docs' six steps.

01

What does the cap count?

✓ Model tokens at each model's list price	TOKENS
✓ Web searches: $10 USD per 1,000	SEARCH
✓ Running time: $0.08 USD per hour	TIME
! List cost, not your contracted price	NOTE
! No budget at creation, none later	LIMIT

Figures are the docs' as of Nov. 27, 2026

02

What happens at the cap?

Docs: Session budgets

Cap is whole US cents

Checked	Idle	Settle only	Resume
Between model requests	stop_reason: budget_reached	Events accepted	Change or remove the cap
OVERSHOOT	PAUSED	400 ON MESSAGES	ONE-WAY
The request that crosses the cap finishes: a 50-cent cap can pause at 53 cents	History and sandbox are kept. Multiagent sessions share one cap	Confirmations, tool results, and interrupts. A user.message is rejected with 400	Any new cap must exceed consumed list cost. Removing it is one-way

03

The stop sequence, in six steps

Interrupt first — STEP 1

Send user.interrupt by itself. A running tool call can finish first, it ends no workflow run, and it does not undo what was done.

Stop what starts it — STEPS 2 TO 4

Pausing a deployment still allows manual runs; archive it to end the schedule. Rotate any exposed key, and ask the agent to stop workflow runs.

Wait for idle — STEP 5

A running session cannot be archived. Wait for it to go idle.

Archive, then delete — STEP 6

Archive to keep the history. Delete only if you do not need it.

READING RULE

- Figures and steps: the docs
- Grouping and tiles: Luminity's
- Prices can change: re-verify
- Informational, not compliance advice

© 2026 Luminity Digital, Inc. | Independent analysis of Anthropic public documentation; not sponsored or endorsed by Anthropic.

Source: Claude Platform Docs: budgets, secure-deployment · platform.claude.com/docs/en/managed-agents/ · beta · verified Nov. 27, 2026

Anthropic and Claude are trademarks of Anthropic PBC.

Watch sessions, review transcripts, and know what is kept

Webhooks announce events and can be missed, so also list sessions on a schedule. A reviewer you run reads transcripts offline.

01

What do you watch?

✓ session.status_idled	IDLE
✓ session.budget_reached	BUDGET
✓ session.status_terminated	ENDED
✓ deployment_run.failed, deployment.paused	DEPLOYMENT
✓ vault_credential.refresh_failed	VAULT

Six events here; the docs list 38 in all

02

What does a webhook delivery guarantee?

Docs: Webhooks (beta)

HTTPS on port 443 only

Retries
Up to three attempts

DELIVERY

Jittered backoff of 5 to 120 seconds. After the last attempt the event is dropped, with no signal

Not durable
No backfill

RECONCILE

Events while unsubscribed or disabled are not replayed. List or fetch to catch up

Out of order
Dedupe on event.id

ORDER

Order is not guaranteed. The event.id is the same across retries

Verify
Signature, 5 minutes

SECURITY

SDK unwrap() errors on an invalid signature or a payload over 5 minutes old

03

What is kept, who reviews, and what you own

Offline reviewer — YOU RUN IT

Optional. A job you run after the fact, separate from the sessions it reviews. Claude flags behavior beyond what the agent was granted.

What is kept — RETENTION

History stays until you delete the session; copy events first. Sandbox state: 30 days after it is created. Session files go with it; uploads, memory stores, vaults do not.

Eligibility — ZDR / HIPAA

Not HIPAA eligible. Not Legacy ZDR eligible, except self-hosted (you store transcripts). Even under ZDR or HIPAA, flagged inputs and outputs may be kept up to 2 years.

Self-hosted: yours — SHARED

You own the image, egress limits, service key, and session-content retention. Anthropic secures multitenant isolation, not tools inside your sandbox.

READING RULE

- Events and retention: the docs
- Grouping and tiles: Luminity's
- Docs verified Nov. 27, 2026
- Informational, not compliance advice

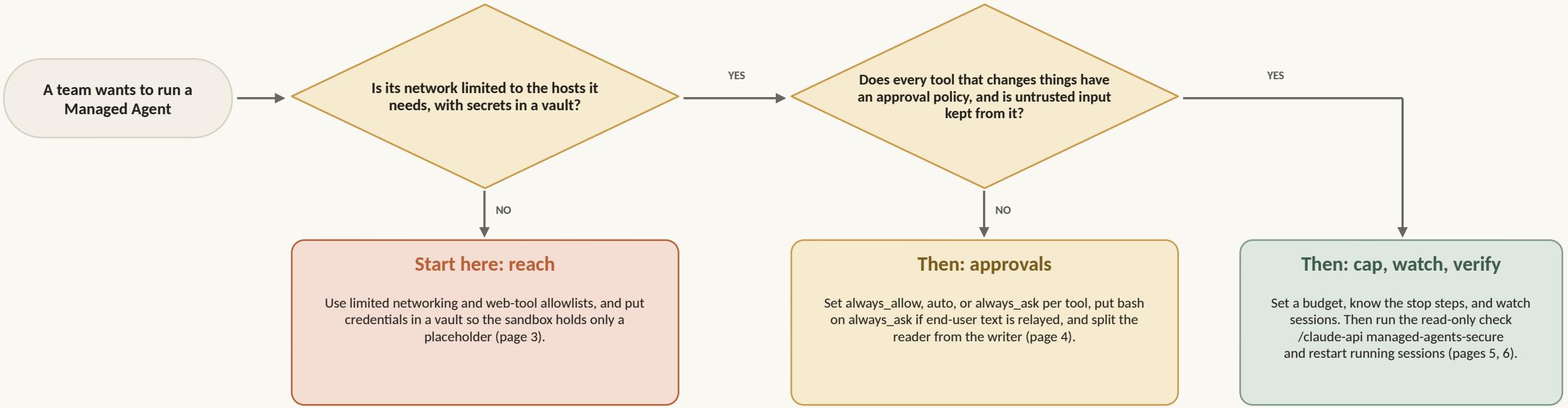
© 2026 Luminity Digital, Inc. | Independent analysis of Anthropic public documentation; not sponsored or endorsed by Anthropic.

Source: Claude Platform Docs: webhooks, secure-deployment-monitoring, self-hosted-sandboxes-security, api-and-data-retention · platform.claude.com/docs/en/ · beta · verified Nov. 27, 2026

Anthropic and Claude are trademarks of Anthropic PBC.

Where do you start?

A path for a first move. The checks come from the docs; the path and wording are Luminity's. Informational, not legal or compliance advice.



READING RULE

- The docs anchor; the path is Luminity's
- Beta: re-verify before relying
- Informational, not compliance advice
- Docs verified Nov. 27, 2026

© 2026 Luminity Digital, Inc. | Independent analysis of Anthropic public documentation; not sponsored or endorsed by Anthropic.

Source: Claude Platform Docs: secure-deployment, permission-policies · platform.claude.com/docs/en/managed-agents/ · beta · verified Nov. 27, 2026

Anthropic and Claude are trademarks of Anthropic PBC.