THE AI TOOLCHAIN

NO. 033

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED SEPTEMBER 21, 2026 · EVERY DAY

The daily firehose — everything the toolchain shipped today, already filtered.

▸ // HOW THIS ISSUE IS MADE SHOW

VIEW

ISSUE VIEW

scroll



SCROLL | EXPAND IN PLACE | PREVIEW PANE

INFORMATION DENSITY

everything



BRIEFED | EVERYTHING

Every tool in the issue, in full.

all 156 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 29 tools

▾ // THE BRIEF

HIDE SAME ISSUE, SAME PROMPT, TWO WRITERS: Opus 5 | GPT-5.5

What stands out across today's releases, grouped by what it lets you do. Every tool named links to its entry below.

Three unrelated things stand out. Braintrust now ingests sandboxed agent eval runs, traces and scores from Harbor without you wiring the

vim ? help

export yourself. compressed-tensors adds MXFP4, NVFP4 and MXFP8 plus a programmatic ModelCompressor API, so a quantized checkpoint is a script step. And Snowflake's `agent:compact` summarizes long agent conversations instead of letting them overrun the context window.

EVALUATE

Get agent eval runs from a sandbox into your scoring project without writing an exporter

Braintrust's native Harbor plugin syncs sandboxed agent runs, traces and scores straight into projects, and it also traces coding sessions for Grok and Google Antigravity — so the eval record covers the agent sessions people actually run, not just the harness. Agent Safe Pipeline's hosted shadow-mode testing and GitHub Actions gate action let you see what a policy would have blocked before you turn it on in a PR check.

Braintrust · agent-safe-pipeline

DEPLOY

Ship a low-bit checkpoint from a script instead of a bespoke export pipeline

compressed-tensors adds MXFP4, NVFP4 and MXFP8 formats, conversion between quantization schemes, and a ModelCompressor API for compressing and saving models programmatically — so requantizing a model for new hardware is code, not a manual re-run. AutoRound 0.15 fills the gaps between the usual 4- and 8-bit stops with 5/6/7-bit weight-only support for vLLM and Transformers, and cuts MoE tuning time 3x.

compressed-tensors · AutoRound

OPERATE

Stop long agent conversations from silently blowing the context window

Cortex Agents' new `agent:compact` endpoint summarizes a running conversation so it stays inside model context limits and costs fewer tokens per turn — the alternative was truncating history yourself and hoping the agent kept the thread. xalgorix attacks the other half of the same bill with per-call token attribution and real provider cache telemetry, so you can see which agent step is expensive rather than inferring it from a monthly total.

Snowflake Cortex AI · xalgorix

Does **Opus 5** read better? 👍 👎

OFFENSIVE SECURITY

◆ Exploitation & C2

xalgorix

Feedback ⓘ v4.6.84 NOTES ↗ RANK

The xalgorix platform runs AI pentesting agents for reconnaissance, vulnerability detection, and exploitation workflows.

> xalgorix v4.6.83–v4.6.84 add fine-grained cost and performance observability for LLM-driven pentesting agents, including per-call token attribution, real provider cache telemetry, skill deduplication, smarter retry handling, and live-tunable agent pacing.

└──▷ WHAT SHIPPED · 6 FEATURES most completely described first ? what's the number?

01 **Live-tunable agent iteration delay** NEW 93

Adds `XALGORIX_ITERATION_DELAY` to pace agent reasoning turns and conserve provider rate limits. Changes made in the Web Settings UI apply immediately to active scans and propagate recursively to running child specialist subagents without a server restart.

Throttle agent reasoning turns to stay within a provider's rolling-window rate limit when launching xalgorix.

```
$ XALGORIX_ITERATION_DELAY=10 xalgorix
```

— Exact env var with runnable example and live-propagation behaviour. v4.6.83

02 **Per-agent skill deduplication** NEW 78

Adds per-agent skill deduplication for `read_skill` and `list_skills` with SHA-256 hash tracking, canonical alias resolution, and re-injection after context compaction or sliding-window pruning.

— Names the exact functions and hashing mechanism involved. v4.6.84

03 **Non-retryable provider error pruning** IMPROVED 65

Adds error handling that immediately aborts on deterministic HTTP 422 and 405 responses, while preserving retries for transient 429 and 5xx responses.

— Names specific status codes and the retry/abort logic. v4.6.84

04 **Multi-dimensional LLM token attribution tracking** NEW 60

Adds a token attribution tracker covering 20+ dimensions per LLM call across coordinator, specialists, and verifier roles, with aggregated post-scan summaries.

— Describes scope and dimensions but no surface to invoke directly. v4.6.84

THINNER COVERAGE BELOW

05 **Provider cache usage telemetry** IMPROVED 55

Adds provider cache telemetry with accurate parsing of MiniMax/OpenAI, Anthropic, and Gemini cache usage and prompt caching support, replacing previously hallucinated estimates.

— Names providers and the fix but no user-facing control. v4.6.84

06 **Engine noise suppression for planner and curl hooks** IMPROVED 45

Suppresses duplicate unchanged plan briefs in `hookPlanner` and consecutive browser/curl warnings in `hookCurlPreference` to reduce engine noise.

— Names the two hooks but describes only a minor cosmetic change. v4.6.84

v4.6.84 NOTES ↗ LATEST STABLE

xalgorix v4.6.84 adds multi-dimensional LLM token attribution tracking and provider cache telemetry across MiniMax/OpenAI, Anthropic, and Gemini.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v4.6.84
https://github.com/xalgorix/xalgorix.git
# already have the repo? check out this version:
$ git checkout v4.6.84
```

Adds a multi-dimensional token attribution tracker covering 20+ dimensions per LLM call across coordinator, specialists, and verifier, with aggregated post-scan summaries.

Adds provider cache telemetry with accurate parsing of MiniMax/OpenAI, Anthropic, and Gemini cache usage and prompt caching support, replacing hallucinated estimates.

Adds per-agent skill deduplication for `read_skill` and `list_skills` with SHA-256 hash tracking, canonical alias resolution, and re-injection after context compaction or sliding-window pruning.

Adds non-retryable provider error pruning that immediately aborts on HTTP 422 and 405 responses and other deterministic errors, while preserving retries for transient 429 and 5xx responses.

Suppresses duplicate unchanged plan briefs in `hookPlanner` and consecutive browser/curl warnings in `hookCurlPreference` to reduce engine noise.

v4.6.83 NOTES ↗ STABLE

xalgorix v4.6.83 adds `XALGORIX_ITERATION_DELAY` to pace agent reasoning turns and conserve provider rate limits.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v4.6.83
https://github.com/xalgorix/xalgorix.git
# already have the repo? check out this version:
$ git checkout v4.6.83
```

└──▷ TRY IT

Throttle agent reasoning turns to stay within a provider's rolling-window rate limit when launching xalgorix.

```
$ XALGORIX_ITERATION_DELAY=10 xalgorix
```

Applies changes to `XALGORIX_ITERATION_DELAY` made in the Web Settings UI immediately to active scans and propagates the new value recursively to running child specialist subagents without a server restart.

WAS THIS USEFUL? 👍 👎

BUILD

◆ AI Coding Agents

Superset

Feedback ⓘ desktop-v1.30.1 NOTES ↗ RANK

Superset is an agentic IDE to orchestrate 100+ coding agents in parallel. Run any agent with your own subscription.

> Superset's biggest release this window is the launch of Superset Mobile for iPhone, letting users monitor and interact with agent terminal sessions and review pull requests from their phone; the desktop app also gained mobile setup/connection improvements, project-qualified pinned sidebar rows, and page-usage auditing in the 10x skill.

└──▷ **WHAT SHIPPED · 6 FEATURES** most completely described first ? what's the number?

01 Terminal agent monitoring and interaction on mobile NEW

72

Enables monitoring and interacting with terminal agent sessions (Claude Code, Codex, and others) from a mobile device — type, dictate, or attach a photo to unblock a running agent.

Unblock a stalled agent session while away from your desk — open the workspace on your iPhone and send the one-line answer it is waiting for.

📍 1. On your iPhone, open Superset Mobile and sign in with your Superset account. 2. Tap the connected machine that is running your agent. 3. Open the workspace and terminal session where the agent is waiting. 4. Type, dictate, or attach a photo with your response to resume the agent.

— Includes step-by-step usage example for unblocking an agent launch-20260921-281f6a32

02 Superset Mobile launch for iPhone NEW

62

Launches Superset Mobile for iPhone (requires iOS 26 and Superset Pro), giving access to existing workspaces and agent sessions from a connected desktop machine.

— Names OS/Pro requirement but no setup command launch-20260921-281f6a32

THINNER COVERAGE BELOW

03 Pull-request review in mobile app NEW

52

Adds pull-request review to the mobile app, with syntax-highlighted diffs viewable file by file and the ability to send agent feedback from the same workspace.

— Describes diff view and feedback flow, no exact UI path launch-20260921-281f6a32

04 **Pro mobile setup and connection status in desktop app**

IMPROVED 43

Adds a Pro setup checklist and gated mobile settings to the desktop app, simplifies mobile setup with inline remote access configuration, and distinguishes a never-connected host from an offline one for clearer connection status feedback.

— Names the checklist and settings but no exact config paths desktop-v1.30.1

05 **Project-qualified pinned sidebar rows** IMPROVED 33

Qualifies pinned sidebar rows with their project name so identically named items are distinguishable.

— Clear behavior described but no navigation detail desktop-v1.30.1

06 **Page-usage auditing in 10x skill** NEW 17

Adds page-usage auditing to the 10x skill.

— Bare mention with no mechanism or scope given desktop-v1.30.1

desktop-v1.30.1 NOTES ↗ LATEST STABLE

Superset desktop v1.30.1 adds a Pro setup checklist, mobile connection status, and project-qualified pinned sidebar rows.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch desktop-v1.30.1 https://github.com/superset-sh/superset.git
# already have the repo? check out this version:
$ git checkout desktop-v1.30.1
```

Adds a Pro setup checklist and gated mobile settings to the desktop app.

Qualifies pinned sidebar rows with their project name so identically named items are distinguishable.

Distinguishes a never-connected host from an offline one in the mobile app, giving clearer connection status feedback.

Simplifies mobile setup with inline remote access configuration in the desktop app.

Adds page-usage auditing to the 10x skill.

announced · Sep 21, 2026 BLOG ↗

Superset Mobile (iOS) lets you monitor agent terminals, send prompts, and review syntax-highlighted diffs from your iPhone

└──▷ HOW TO FIND IT

Unblock a stalled agent session while away from your desk — open the workspace on your iPhone and send the one-line answer it is waiting for.

> 📍 1. On your iPhone, open Superset Mobile and sign in with your Superset account. 2. Tap the connected machine that is running your agent. 3. Open the workspace and terminal session where the agent is waiting. 4. Type, dictate, or attach a photo with your response to resume the agent.

Launches Superset Mobile for iPhone (requires iOS 26 and Superset Pro), giving access to existing workspaces and agent sessions from a connected desktop machine.

Enables monitoring and interacting with terminal agent sessions (Claude Code, Codex, and others) from a mobile device — type, dictate, or attach a photo to unblock a running agent.

Adds pull-request review to the mobile app, with syntax-highlighted diffs viewable file by file and the ability to send agent feedback from the same workspace.

WAS THIS USEFUL? 👍 👎

OpenAI Codex CLI

Feedback ⓘ rust-v0.156.0-alpha.13 NOTES ↗

CODE ↗ RANK

OpenAI Codex CLI runs an agent in the terminal that reads, changes, and tests code in local repositories.

> Codex CLI's TUI got a major interaction overhaul this window — new transcript search, mouse-driven text selection and copying, a `/tui` mode switch, a warnings viewer, and persistent quota indicators — alongside two breaking config/keybinding changes.

└──▷ **WHAT SHIPPED · 15 FEATURES** most completely described first ? what's the number?

01 /tui command and fullscreen transcript config NEW

85

Adds a `/tui` slash command inside a session to pick between Scrollback and Fullscreen terminal UI modes, persisting the choice to `tui.fullscreen_transcript` (also exposed as the `tui_fullscreen_transcript` config key) in the user config file; the change takes effect on next restart.

— Exact command and config key with persistence behavior given. rust-v0.156.0-alpha.11 rust-v0.156.0-alpha.10

02 Mouse and keyboard text selection in transcript and composer

NEW 80

Adds right-click copying for selected text in the transcript and composer panes (selection retained after failed or unconfirmed copies), plus mouse selection, word/line selection, drag autoscrolling, and keyboard selection with Ctrl+Space in the transcript viewer, supporting copying and opening links. Plain left-click now opens transcript links, bare URLs get Markdown link styling, mouse selection and editing was added to the fullscreen composer, and the interactive transcript viewer was integrated into the alternate-screen TUI with scrolling, selection, copying, links, and older-history pagination.

— Names exact keys and interactions but no limits or numbers. rust-v0.156.0-alpha.10 rust-v0.156.0-alpha.11 rust-v0.156.0-alpha.12

03 Transcript search NEW

75

Adds incremental, case-insensitive transcript search in the TUI, triggered with `F3` or `/` in the transcript pager, with match navigation in both directions.

— Gives exact trigger keys and search behavior. rust-v0.156.0-alpha.10

04 tui_whimsy config key replaced by tui_effects

BREAKING

65

Replaces the `tui_whimsy` config key with `tui_effects` for independent control of TUI visual effects; existing configurations referencing `tui_whimsy` no longer apply.

— Both exact config keys named with migration impact. rust-v0.156.0-alpha.10

05 **Persistent quota usage indicator in composer** NEW 60

Shows the most constrained quota window in the composer hint row, including remaining usage, reset time, and a `/status` link — stays visible while typing or while a turn is running.

— Names /status link and hint-row behavior but no numeric limits. rust-v0.156.0-alpha.13

06 **Task deletion keybinding changed to Backspace** BREAKING 60

The default key binding for task deletion in the agent command center is changed from `Delete` to `Backspace`.

— Exact before/after key named, clear migration note. rust-v0.156.0-alpha.11

THINNER COVERAGE BELOW

07 **Per-activity detail controls in transcript** NEW 55

Adds per-activity detail controls and `F4` keyboard focus for inspecting individual activity groups in the transcript.

— Names the key but little on what detail is shown. rust-v0.156.0-alpha.10

08 **Subagent MCP elicitation requests** NEW 40

Allows subagents to request MCP elicitation input.

— Names MCP elicitation but no usage detail. rust-v0.156.0-alpha.10

09 **Warnings viewer in TUI** NEW 35

Adds a dedicated warnings viewer and warning footer to the TUI.

— Only a bare description of the new viewer. rust-v0.156.0-alpha.10

10 **Status filter tabs in agent command center** NEW 35

Adds status filter tabs to the agent command center.

— Names the UI addition but no detail on filter options. rust-v0.156.0-alpha.10

11 **Uniform transcript list spacing** IMPROVED 35

Makes transcript list spacing uniform after streaming completes, separating multi-row sibling items and keeping single-line lists compact with reflow on resize.

— Describes layout behavior but no user-facing control. rust-v0.156.0-alpha.12

12 **Mermaid stadium node rendering** NEW 30

Adds support for stadium nodes in Mermaid terminal diagram rendering.

— Names the node type but nothing else. rust-v0.156.0-alpha.10

13 **Mouse navigation in usage view** NEW 25

Enables mouse navigation in the TUI usage view.

— Bare one-line description with no mechanism. rust-v0.156.0-alpha.10

14 **System clock preference in timestamps** IMPROVED 25

Honors system clock preference in TUI completion timestamps.

— Minimal description of a display tweak. rust-v0.156.0-alpha.10

15 **Animated logo on fresh conversations** NEW 25

Adds an animated Codex logo to fresh conversations and onboarding.

— Cosmetic addition with no further detail. rust-v0.156.0-alpha.10

└──▷ BREAKING ON UPGRADE

! The default key binding for task deletion in the agent command center is changed from `Delete` to `Backspace`.

! The `tui_whimsy` config key is replaced by `tui_effects`; existing configurations referencing `tui_whimsy` will no longer apply.

rust-v0.156.0-alpha.13 NOTES ↗ CODE ↗ PRE-RELEASE

Codex CLI v0.156.0-alpha.13 adds local background server identification in `/status` and persistent quota warnings in the TUI.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch rust-v0.156.0-alpha.13
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.156.0-alpha.13
```

Shows the most constrained quota window in the composer hint row, including remaining usage, reset time, and `/status` link — kept visible while typing or running a turn.

rust-v0.156.0-alpha.12 NOTES ↗ CODE ↗ PRE-RELEASE

Codex CLI adds right-click copying in the transcript and composer, plus accurate terminal color rendering in activity charts.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch rust-v0.156.0-alpha.12
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.156.0-alpha.12
```

Adds right-click copying for selected text in the transcript and composer panes, with selection retained after failed or unconfirmed copies.

Makes transcript list spacing uniform after streaming completes, separating multi-row sibling items and keeping single-line lists compact with reflow on resize.

rust-v0.156.0-alpha.11 NOTES ↗ CODE ↗ PRE-RELEASE

Adds `/tui` slash command to switch between Scrollback and Fullscreen terminal UI modes and saves the preference to config.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch rust-v0.156.0-alpha.11
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.156.0-alpha.11
```

Adds `/tui` slash command inside a session to pick between Scrollback and Fullscreen terminal UI modes, persisting the choice to `tui.fullscreen_transcript` in the user config file (takes effect on next restart).

Enables plain (unmodified) left-click to open transcript links, and applies Markdown link styling to bare URLs in the transcript.

└──▷ BREAKING ON UPGRADE

! The default key binding for task deletion in the agent command center is changed from `Delete` to `Backspace`.

rust-v0.156.0-alpha.10 NOTES ↗ CODE ↗ PRE-RELEASE

Codex CLI TUI gains transcript search, mouse selection, activity detail controls, warnings viewer, and agent command center improvements.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch rust-v0.156.0-alpha.10
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.156.0-alpha.10
```

Adds `tui_fullscreen_transcript` config key to control fullscreen transcript display.

Adds transcript search (incremental, case-insensitive) in the TUI, triggered with `F3` or `/` in the transcript pager, with match navigation in both directions.

Adds per-activity detail controls and `F4` keyboard focus for inspecting individual activity groups in the transcript.

Adds mouse selection, word/line selection, drag autoscrolling, and keyboard selection with `Ctrl+Space` in the transcript viewer, with support for copying selected text and opening links.

Adds a dedicated warnings viewer and warning footer to the TUI.

+9 more

└──▷ BREAKING ON UPGRADE

! The `tui_whimsy` config key is replaced by `tui_effects`; existing configurations referencing `tui_whimsy` will no longer apply.

WAS THIS USEFUL?  

Letta Code

Feedback ⓘ v0.32.15 NOTES ↗ RANK

Letta Code is a terminal coding agent with persistent memory and identity across development sessions.

> Letta Code v0.32.15 adds a workflow tool that wraps the Agent SDK query() method and forwards structured output schemas through the protocol layer.

└──▷ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 **Workflow tool wrapping Agent SDK query()** NEW 55

A new workflow tool acts as a thin wrapper around the Agent SDK query() method, enabling agents to invoke structured workflows programmatically.

— Names the wrapped method but no usage example or config. v0.32.15

02 **Structured output schemas forwarded through protocol layer**

IMPROVED 45

The protocol layer now forwards structured output schemas, allowing downstream consumers to receive and enforce typed response shapes.

— Describes mechanism but no concrete schema or endpoint named. v0.32.15

v0.32.15 NOTES ↗ LATEST STABLE

Letta Code v0.32.15 adds a workflow tool wrapping Agent SDK query() and forwards structured output schemas through the protocol layer.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.32.15 https://github.com/letta-ai/letta-code.git
# already have the repo? check out this version:
$ git checkout v0.32.15
```

Adds a workflow tool as a thin wrapper around the Agent SDK query() method, enabling agents to invoke structured workflows programmatically.

Forwards structured output schemas through the protocol layer, allowing downstream consumers to receive and enforce typed response shapes.

WAS THIS USEFUL? 👍 👎

Worktrunk

Feedback ⓘ v0.79.0 NOTES ↗ RANK

Worktrunk manages Git worktrees from the command line for parallel AI coding-agent workflows, with hooks, status listings, and LLM-written commits.

> Worktrunk's v0.79.0 release changes where `wt switch --no-cd -x` executes programs, adds config-file migration support, tweaks picker preview visuals, and stops rebases from dragging along stacked branches.

└──▷ **WHAT SHIPPED · 4 FEATURES** most completely described first ? what's the number?

01 wt switch --no-cd runs programs in worktree directory

80

BREAKING

`wt switch --no-cd -x <program>` now runs the executed program inside the target worktree's directory rather than the invoking directory, enabling one-liners like `wt switch feature --no-cd -x code -- .` to open an editor on the worktree without moving your terminal. This is a breaking change: a program that previously relied on the invoking directory must now pass an explicit `cd` or path.

Launch an editor on a worktree from your current terminal without changing your shell's working directory.

```
$ wt switch feature --no-cd -x code -- .
```

— Named flag, clear before/after, and a runnable example. v0.79.0

02 Config file migration in wt config update IMPROVED

70

`wt config update --output` now accepts the config file being migrated, moving `approved-commands` to `approvals.toml` before rewriting.

— Names exact flag and config keys but no runnable example. v0.79.0

03 rebase.updateRefs no longer drags stacked branches BREAKING

65

`rebase.updateRefs` no longer moves other local branches stacked in the rebased range when `wt step rebase` or `wt merge` runs, changing prior behavior that shifted those refs.

— Named config option and commands but no example given. v0.79.0

THINNER COVERAGE BELOW

04 **Picker preview tab label indicators** IMPROVED 30

Picker preview tab labels now show dim when a tab has no content for the selected row and underlined when active, making content availability visible at a glance.

— UI tweak described but no interactive surface named. v0.79.0

└──▷ **BREAKING ON UPGRADE**

! Programs invoked via `wt switch --no-cd -x <program>` now start in the worktree directory instead of the invoking directory; a program that previously relied on the invoking directory must now pass an explicit `cd` or path.

! `rebase.updateRefs` no longer moves other local branches stacked in the rebased range when `wt step rebase` or `wt merge` runs.

v0.79.0 NOTES ↗ LATEST STABLE

wt switch `--no-cd` now runs exec'd programs in the worktree, and the `/wt-switch-create` Claude Code plugin no longer stalls at path-entry dialogs.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch v0.79.0 https://github.com/max-sixty/worktrunk.git
# already have the repo? check out this version:
$ git checkout v0.79.0
```

└──▷ **TRY IT**

Launch an editor on a worktree from your current terminal without changing your shell's working directory.

```
$ wt switch feature --no-cd -x code -- .
```

Changes `wt switch --no-cd -x <program>` so the executed program runs inside the target worktree's directory rather than the invoking directory, enabling one-liners like `wt switch feature --no-cd -x code -- .` to open an editor on the worktree without moving your terminal.

Allows `wt config update --output` to accept the config file being migrated, moving `approved-commands` to `approvals.toml` before rewriting.

Improves picker preview tab labels to show dim when a tab has no content for the selected row and underlined when active, making content availability visible at a glance.

└──▷ **BREAKING ON UPGRADE**

- ! Programs invoked via `wt switch --no-cd -x <program>` now start in the worktree directory instead of the invoking directory; a program that previously relied on the invoking directory must now pass an explicit `cd` or path.
- ! `rebase.updateRefs` no longer moves other local branches stacked in the rebased range when `wt step rebase` or `wt merge` runs.

WAS THIS USEFUL? 👍 👎

Graphify Labs Graphify

Feedback ⓘ v0.9.65 NOTES ↗

RANK

Graphify converts code, docs, SQL schemas, configs, and PDFs into a queryable knowledge graph with explainable relationships.

> Graphify v0.9.65 extends its code graph with interface and protocol method resolution for Go and Swift, alongside a Pillow dependency bump that changes install behavior for the svg/all extras on old Linux hosts.

└──▷ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 **Pillow >=12.3.0 requirement for svg/all extras** BREAKING

70

The `svg` / `all` extras now require Pillow `>=12.3.0`. Pillow 12.3.0 dropped its glibc<2.27 cp310 Linux wheel, so Python 3.10 hosts on very old glibc installing the `svg` or `all` extra will build Pillow from source (sdist) instead of installing a pre-built wheel.

— Names exact extras, version threshold, and platform impact v0.9.65

02 **Interface and protocol method resolution for Go and Swift**

NEW

60

Go interface method requirements written as `type Foo interface { Bar() }` are now added as nodes attached to the interface, enabling call resolution to interface methods. Swift protocol method requirements written as `protocol P { func f() }` are similarly added as nodes attached to the protocol node, enabling call resolution to protocol methods.

— Names exact syntax nodes but no broader mechanism or limits given v0.9.65

└──▷ BREAKING ON UPGRADE

! The `svg` / `all` extras now require Pillow `>=12.3.0`; Pillow 12.3.0 dropped its glibc<2.27 cp310 Linux wheel, so Python 3.10 hosts on very old glibc with the `svg` or `all` extra will build Pillow from source (sdist) instead of installing a pre-built wheel.

v0.9.65 NOTES ↗ LATEST STABLE

Graphify v0.9.65 adds Go interface and Swift protocol method resolution to the code graph.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.9.65 https://github.com/Graphify-Labs/graphify.git
# already have the repo? check out this version:
$ git checkout v0.9.65
```

Adds Go interface method requirements (`type Foo interface { Bar() }`) as nodes attached to the interface, enabling call resolution to interface methods.

Adds Swift protocol method requirements (`protocol P { func f() }`) as nodes attached to the protocol node, enabling call resolution to protocol methods.

└──▷ **BREAKING ON UPGRADE**

! The `svg` / `all` extras now require Pillow `>=12.3.0` ; Pillow 12.3.0 dropped its glibc<2.27 cp310 Linux wheel, so Python 3.10 hosts on very old glibc with the `svg` or `all` extra will build Pillow from source (sdist) instead of installing a pre-built wheel.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Code

Feedback ⓘ v0.24.3 NOTES ↗ RANK

An open-source AI coding agent that lives in your terminal.

> Qwen Code added a JDBC-backed persistence layer for Runtime sessions, a bwrap-based Linux sandbox foundation, and new trajectory/transcript-replay visualizations in Web Shell, alongside Chrome extension packaging, Live Voice audio improvements, and workflow retry via `serve`.

└──▷ **WHAT SHIPPED · 15 FEATURES** most completely described first ? what's the number?

01 Trajectory visualization and transcript replay in Web Shell

NEW

80

Web Shell gains an opt-in trajectory table tab showing per-turn durations, time to first token, token counts, and tool timing; paged transcript events are now projected into a trajectory, and the `acp-bridge` surfaces request and tool timing on paged transcript replay.

— Multiple named metrics and modules with screenshots v0.24.3 v0.24.2

02 Structured shell execution results in Web Shell IMPROVED

70

Web Shell now presents structured shell execution results with command, output, execution details, elapsed time, and scrollable content.

— Lists concrete UI fields with supporting screenshots v0.24.3

03 Chrome extension packaging and release workflow NEW

70

Adds Chrome Web Store packaging and a release workflow that can build, scan, upload, and optionally publish the Chrome extension during stable releases or manual dispatch.

— Describes workflow stages but no exact command v0.24.3

04 Live Voice audio and session improvements IMPROVED

70

Live Voice's microphone capture now uses an AudioWorklet, with the old ScriptProcessorNode kept as fallback, reducing main-thread work while preserving level metering and call gating; a browser Host can now answer a screen request from a shared screen; users can select the Live Voice model and voice from the setup card in the web-shell; and the web-shell shows microphone level during a browser Live Voice call.

— Groups four Live Voice increments with named mechanisms v0.24.3 v0.24.2

05 **bwrap-based Linux sandbox execution foundation** NEW

65

Adds an internal `bwrap -based` Linux sandbox execution foundation with structured process launching, supervision, and confined worker support.

— Names the sandbox tool but no user-facing entry point v0.24.2

06 **JDBC persistence for Runtime bindings and sessions** NEW

60

Adds JDBC persistence for Runtime bindings and sessions, enabling shared state across broker processes and restart-safe ownership.

— Names mechanism but no user-facing command shown v0.24.3

07 **Retry and rerun workflow runs via serve** NEW

60

Enables retrying and rerunning workflow runs from persisted history via `serve`.

— Names the `serve` command as a starting point v0.24.2

THINNER COVERAGE BELOW

08 **argsUnavailable retry signaling for history entries** IMPROVED

55

History entries that cannot be retried now expose `argsUnavailable` so clients can accurately reflect retry limitations.

— Named field but no usage example given v0.24.3

09 **Native title bar on macOS desktop** IMPROVED

55

macOS desktop now uses an integrated native title bar with content painted behind traffic-light controls and a dedicated drag region.

— Describes UI change but nothing for reader to configure v0.24.3

10 **Pipe-based shell backend for one-shot headless execution** IMPROVED

55

One-shot headless execution now defaults to a pipe-based shell backend and optimizes relaunch behavior for lower overhead.

— Names the backend but no flag or command given v0.24.3

11 **Versioned private protocol foundation for Hosted Harness** NEW

55

Defines a versioned private protocol foundation for Hosted Harness, including capability envelope, boot fence, and request validation.

— Internal foundation with named parts, not user actionable v0.24.3

12 **Context file token cost attribution in CLI output** NEW

55

Attributes each context file to its extension and reports its token cost in `cli` output.

— Names `cli` output surface but limited mechanism detail v0.24.2

13 **Core session and context handling improvements** NEW

55

In `core`, conditional extension rules and caps on context warnings were added, background turns can now carry a message from another session (cross-session messaging), and prompt cache is preserved for deferred tools.

— Groups three thin core increments, no exact config keys v0.24.2

14 **Web Shell artifact and workspace UI updates** IMPROVED

55

Web-shell now exposes slash-command exports as artifacts, opens link artifacts in a new page, and restores the remote workspace add flow.

— Groups three thin UI fixes with clear navigation points v0.24.2

15 **Concurrent browser-use sessions sharing a Chrome profile** NEW

40

Supports concurrent browser-use sessions sharing a single Chrome profile.

— Bare capability statement, no mechanism detail v0.24.2

v0.24.3 NOTES ↗ LATEST STABLE

Qwen Code v0.24.3 adds JDBC session persistence, Web Shell trajectory metrics, Chrome extension packaging, and `argsUnavailable` retry signaling.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.24.3 https://github.com/QwenLM/qwen-code.git
# already have the repo? check out this version:
$ git checkout v0.24.3
```

Adds JDBC persistence for Runtime bindings and sessions, enabling shared state across broker processes and restart-safe ownership.

History entries that cannot be retried now expose `argsUnavailable` so clients can accurately reflect retry limitations.

Adds an opt-in trajectory table tab in Web Shell showing per-turn durations, time to first token, token counts, and tool timing.

Web Shell now presents structured shell execution results with command, output, execution details, elapsed time, and scrollable content.

Adds Chrome Web Store packaging and a release workflow that can build, scan, upload, and optionally publish the Chrome extension during stable releases or manual dispatch.

+5 more

v0.24.2 NOTES ↗ STABLE

Qwen Code v0.24.2 adds a bwrap-based Linux sandbox, concurrent browser-use sessions, workflow retry from history, and cross-session messaging.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.24.2 https://github.com/QwenLM/qwen-code.git
# already have the repo? check out this version:
$ git checkout v0.24.2
```

Adds an internal `bwrap` `-based` Linux sandbox execution foundation with structured process launching, supervision, and confined worker support.

Supports concurrent browser-use sessions sharing a single Chrome profile.

Enables retrying and rerunning workflow runs from persisted history via `serve`.

Attributes each context file to its extension and reports its token cost in `cli` output.

Adds conditional extension rules and caps context warnings in `core`.

+9 more

WAS THIS USEFUL? 👍 👎

Red Hat ripwire

Feedback ⓘ v0.6.2 NOTES ↗ RANK

ripwire gives coding agents a ranked, deterministic map of a repository — call graph, blast radius, and tests to run — from a CLI or over MCP.

> ripwire shipped an MCP server mode (stdio and socket-listening) exposing 33 verbs, added native Windows x64 builds, expanded language coverage to Kotlin, GDScript, Elixir and Dart, and improved call-site attribution for code outside named functions across 12 languages; it also added a self-contained interactive HTML call-graph viewer and new flags for scoping and capping output.

└──▷ **WHAT SHIPPED · 11 FEATURES** most completely described first ? what's the number?

01 MCP server mode with stdio and socket listener NEW

95

Adds `--mcp --listen=HOST:PORT` mode to bind a socket and serve other processes, with `--mcp-token` and `--allow-remote-edits` flags for access control, alongside a `--mcp` stdio server exposing 33 verbs (17 read, 13 flagship-reflex, 3 span-addressed edit). MCP definitions are now sent once per session instead of in every answer.

— Exact flags, verb counts and behaviour change are all named. product docs

02 Output and context size controls NEW

90

Adds `--token-budget=N` to cap the output bundle size and `--top-k=N` to cap rows on the default map, `--query`, `--format=candidates`, `--recall`, and `--graph-query`. Adds `--signatures-only` to drop auto-bodies from `--for` output, and `--detail=N` to include full bodies for only the top N results.

Cap context bundle size when feeding ripwire output to an LLM agent with a tight token limit.

```
$ ripwire . --for="validate user session" --token-budget=4000 --signatures-only
```

— Four flags named with scope and a runnable example. product docs

03 Call attribution for code outside named functions IMPROVED

90

Adds caller context for call sites that live outside any named function, so `--callers`, `--impact`, and `--test-gate` cover previously invisible call sites (72.83% coverage measured on vue-core). Extends this with a module-scope owner (`t="modscope"`, named `<file-scope>`) per file for such calls, so `--callers`, `--impact`, `--`

`affected`, and `--test-gate` no longer answer short on top-level statements, anonymous callbacks, DI wiring, and framework registration code across all 12 affected indexed languages. `--expand` now returns `bodyless="1" capped="0"` for `<file-scope>` module-scope owners to distinguish them from normal callees when expanded.

— Names response fields, affected verbs, language count and a measured figure. v0.6.2

04 **Self-contained interactive HTML call-graph viewer** NEW

90

Adds `--html[=FILE]` to emit a self-contained, single-file interactive call-graph visualization with no server, CDN, or external assets required. Adds `--color-by=lang|community|cx|churn|tested` to set the initial colour lens, with all five lenses embedded and switchable in the page without a second run.

Generate an interactive, self-contained call-graph HTML file coloured by churn to spot hotspots before a refactor.

```
$ ripwire . --html=graph.html --color-by=churn
```

— Exact flags, five named lenses and a runnable example. product docs

05 **Native Windows build support** IMPROVED

85

Adds native Windows x64 build support compiling with both clang-cl and MSVC `cl.exe`, verified in CI; a later release extends this so both Windows front ends compile and pass the full CI matrix including smoke steps (`--version`, `ctest`, fixture crawl, and the two-run byte-identical contract). A new arm H in `test/osswitchcheck.sh` with pinned `EXEMPT_COUNTS` per row now refuses any new `__builtin_*`, inline asm, or `__attribute__` outside the platform portability seam pair, so a Windows-breaking change fails on every POSIX leg.

— Names compilers, script path and CI checks, but is a build/CI concern. v0.6.2

06 **flow_redundant flag on unseeded slice-flow output** NEW

80

Adds a `flow_redundant="1"` attribute to the `<slice>` root when `--slice-flow=both` runs without an `--at=` seed, disclosing that the unseeded result adds nothing beyond the flat rows rather than silently returning a less-informative answer.

— Named attribute and flag with the exact condition it fires under. v0.6.2

07 **New language support: Kotlin, GDScript, Elixir, Dart** NEW

75

Adds Kotlin (`.kt`) support: classes, objects, companion objects, interfaces, enum classes, extension functions, and call-graph edges shared with Java. Adds GDScript (`.gd`) support: `class_name` , file-scope `func` / `var` , inner classes, constants, enums, signals, and call edges. Adds Elixir native resolution by module, name, and arity. Adds Dart language support, bringing vendored grammars to 24 total.

— Enumerates language constructs but no commands to invoke them. product docs

THINNER COVERAGE BELOW

08 **Compact legend now default** BREAKING 50

The compact legend is now the default output for map results; the previous full-length legend can be restored with `--legend=full` .

— Named flag but only one line of behaviour described. product docs

09 **Directory-scoped change detection with --in=DIR** NEW 50

Adds `--in=DIR` to scope what changed recently analysis to a specific directory.

— Single named flag with no further mechanism given. product docs

10 **file:name selector syntax** NEW 50

Adds a `file:name` selector syntax to prevent definitions from resolving across files.

— Named syntax with a clear one-line purpose. product docs

11 **scip-java index support via --scip** NEW 45

Adds `--scip` to read scip-java indexes.

— Named flag but no detail on scope or behaviour. product docs

└──▷ BREAKING ON UPGRADE

! The compact legend is now the default output; the previous full legend is only restored with `--legend=full` .

v0.6.2 NOTES ↗ LATEST STABLE

ripwire v0.6.2 adds module-scope call ownership across 12 languages, fixing phantom zeros in `--callers` , `--impact` , `--affected` , and `--test-gate` .

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.6.2 https://github.com/redhat-et/ripwire.git
# already have the repo? check out this version:
$ git checkout v0.6.2
```

Adds a module-scope owner (`t="modscope"` , named `<file-scope>`) per file for calls written outside any named function, so `--callers` , `--impact` , `--affected` , and `--test-gate` no longer answer short on top-level statements, anonymous callbacks, DI wiring, and framework registration code — across all 12 affected indexed languages.

Adds `flow_redundant="1"` attribute to the `<slice>` root when `--slice-flow=both` runs without an `--at=` seed, disclosing that the unseeded result adds nothing beyond the flat rows rather than silently returning a less-informative answer.

Adds `--expand` response fields `bodyless="1" capped="0"` for `<file-scope>` module-scope owners, distinguishing them from normal callees when expanded.

Enables native builds with `cl.exe` (MSVC) in addition to clang-cl, so both Windows front ends now compile and pass the full CI matrix including smoke steps (`--version` , `ctest` , fixture crawl, and the two-run byte-identical contract).

Adds arm H to `test/osswitchcheck.sh` with pinned `EXEMPT_COUNTS` per row, refusing any new `__builtin_*` , inline asm, or `__attribute__` outside the platform portability seam pair — so a Windows-breaking change now fails on every POSIX leg.

WAS THIS USEFUL?  

Charm Crush

Feedback ⓘ v0.96.0 NOTES ↗ CODE ↗ RANK

Terminal-based AI coding assistant that integrates LLMs, LSPs, and MCP servers into your development workflow.

> Crush v0.96.0 introduces a full in-TUI theme system with live preview and a built-in Gruvbox Dark theme, alongside real-time Hyper credit balance display during streaming responses.

└──▷ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 **Theme system with live preview** NEW 67

A new Themes dialog, accessible from the command palette via `switch_theme`, lets users preview themes live and switch between them. It ships with a built-in Gruvbox Dark theme and supports creating and editing custom palettes directly inside the TUI.

— Names command and UI path but no config file format for custom palettes v0.96.0

THINNER COVERAGE BELOW

02 **Real-time Hyper credit balance display** IMPROVED 30

Hyper credit balance now updates in real time as responses stream, rather than only refreshing after a prompt completes.

— Explains before/after behaviour but no UI location or mechanism v0.96.0

v0.96.0 NOTES ↗ CODE ↗ LATEST STABLE

Crush v0.96.0 adds a full theme system with live preview and built-in Gruvbox Dark, plus real-time Hyper credit balance display.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.96.0
https://github.com/charmbracelet/crush.git
# already have the repo? check out this version:
$ git checkout v0.96.0
```

Adds a Themes dialog accessible from the command palette (`switch_theme`) with live preview, built-in Gruvbox Dark theme, and support for creating and editing custom palettes inside the TUI.

Displays Hyper credit balance in real time as responses stream, rather than only updating after a prompt completes.

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

LangChain

Feedback ⓘ langchain-typesafe==0.0.1a3 NOTES ↗ RANK

LangChain is an open-source framework that orchestrates applications powered by language models.

> LangChain introduced a new experimental package, langchain-typesafe, for type-safe LLM routing and classification.

└──▷ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 langchain-typesafe package for type-safe LLM routing NEW

60

New alpha package `langchain-typesafe==0.0.1a3` adds `TypeSafeClassifier` for type-safe classification of LLM inputs/outputs, experimental `AutoModeMiddleware` for automatic mode selection in LLM invocation pipelines, and experimental `ModelRouterMiddleware` for routing requests across models within those pipelines. Classifier questions are now invocation-scoped, giving per-call control over classification behavior.

— Names four new classes/behaviors but no usage code or mechanism detail.

langchain-typesafe==0.0.1a3

---

langchain-typesafe==0.0.1a3 NOTES ↗ PRE-RELEASE

langchain-typesafe debuts with TypeSafeClassifier, AutoModeMiddleware, and ModelRouterMiddleware for type-safe LLM routing.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch langchain-typesafe==0.0.1a3
https://github.com/langchain-ai/langchain.git
# already have the repo? check out this version:
$ git checkout langchain-typesafe==0.0.1a3
```

Adds `TypeSafeClassifier` for type-safe classification of LLM inputs/outputs.

Adds experimental `AutoModeMiddleware` for automatic mode selection in LLM invocation pipelines.

Adds experimental `ModelRouterMiddleware` for routing requests across models within LLM invocation pipelines.

Makes classifier questions invocation-scoped, allowing per-call control over classification behavior.

WAS THIS USEFUL? 👍 👎

Stacklok Mecatl

Feedback ⓘ v0.0.39 NOTES ↗ RANK

Mecatl is an open-source agent harness for running the same provider-agnostic agent loop locally, remotely, or on Kubernetes, with durable state, permissions, and gRPC/HTTP-SSE APIs.

> Mecatl v0.0.39 focuses on MCP integration, adding OAuth-based authorization and workspace enrollment to the SDK plus streamlined onboarding in `mecated`, alongside run-ID-addressed run controls, configurable command runners, a unified provider setup UX, and protocol-level prompt cache negotiation.

└──▷ **WHAT SHIPPED · 5 FEATURES** most completely described first ? what's the number?

01 **MCP OAuth onboarding and workspace enrollment** NEW 50

The SDK gains an MCP authorization lifecycle enabling OAuth-based auth flows for MCP connections, and MCP workspace enrollment lets connectors programmatically join a workspace. `mecated` now streamlines direct MCP OAuth onboarding, simplifying the end-to-end OAuth flow for MCP servers.

— Names SDK, mecated and MCP but no flags or endpoints. v0.0.39

02 **Protocol-level prompt cache negotiation** IMPROVED 40

Requests prompt cache through the protocol layer rather than the vendor, making cache negotiation provider-agnostic.

— Explains mechanism shift but no flag or config given. v0.0.39

03 **Run-ID-addressed run controls in SDK** NEW 35

Adds run-ID-addressed controls to the SDK, allowing targeted control of specific agent runs by run ID.

— Describes capability but no method signature or example. v0.0.39

04 **Unified provider setup UX** IMPROVED 35

Completes provider setup UX on unified configuration, enabling full provider onboarding through a single config surface.

— No concrete config key or UI path named. v0.0.39

05 **Configurable command runners** NEW 20

Adds configurable command runners so execution environments can be customized per deployment.

— Bare description with no config keys or details. v0.0.39

v0.0.39 NOTES ↗ LATEST STABLE

Mecatl v0.0.39 adds MCP OAuth onboarding, MCP authorization lifecycle in the SDK, run-ID-addressed controls, MCP workspace enrollment, configurable command runners, and unified provider setup UX.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.0.39 https://github.com/stacklok/mecatl.git
# already have the repo? check out this version:
$ git checkout v0.0.39
```

Adds MCP authorization lifecycle to the SDK, enabling OAuth-based auth flows for MCP connections.

Adds run-ID-addressed controls to the SDK, allowing targeted control of specific agent runs by run ID.

Exposes MCP workspace enrollment through the SDK so connectors can programmatically join a workspace.

Streamlines direct MCP OAuth onboarding in `mecated`, simplifying the end-to-end OAuth flow for MCP servers.

Adds configurable command runners so execution environments can be customized per deployment.

+2 more

WAS THIS USEFUL? 👍 👎

LangChain LangGraph

Feedback ⓘ 1.2.12 NOTES ↗ RANK

Low-level orchestration framework for building, managing, and deploying long-running, stateful AI agents.

> LangGraph 1.2.12 adds a `response_schema` parameter to interrupt(), letting developers declare the expected structure of a human response when pausing graph execution.

└──▷ **WHAT SHIPPED · 1 FEATURE** most completely described first ? what's the number?

01 **response_schema parameter for interrupt()** NEW 50

Adds a `response_schema` parameter to interrupt() to declare the expected structure of a human response when pausing graph execution for human-in-the-loop workflows.

— Names exact parameter and function but no usage example or schema format. 1.2.12

1.2.12 NOTES ↗ LATEST STABLE

LangGraph 1.2.12 adds `response_schema` to interrupt() for structured human-in-the-loop responses.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch 1.2.12 https://github.com/langchain-ai/langgraph.git
# already have the repo? check out this version:
$ git checkout 1.2.12
```

Adds `response_schema` parameter to interrupt() to declare the expected structure of a human response when pausing graph execution.

WAS THIS USEFUL? 👍 👎

LangChain Deep Agents

Feedback ⓘ deepagents==0.7.16

NOTES ↗ RANK

Deep Agents provides an agent harness with planning, filesystem tools, subagents, and context management.

> LangChain Deep Agents shipped a single fix enabling dynamic skill reloading when skill metadata is cleared.

└──▷ **WHAT SHIPPED · 1 FEATURE** most completely described first ? what's the number?

01 **Automatic skill reload on metadata reset** IMPROVED 50

Deep Agents now reloads skills automatically when `skills_metadata` is reset to `None`, enabling dynamic skill refresh at runtime.

– Names exact config key but no broader mechanism detail. deepagents==0.7.16

deepagents==0.7.16 NOTES ↗ LATEST STABLE

deepagents 0.7.16 reloads skills when `skills_metadata` is reset to `None`

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch deepagents==0.7.16
https://github.com/langchain-ai/deepagents.git
# already have the repo? check out this version:
$ git checkout deepagents==0.7.16
```

WAS THIS USEFUL? 👍 👎

◆ MCP Servers & Tooling

gram

Feedback ⓘ server@2.9.0 NOTES ↗ RANK

Gram provides MCP servers and AI-agent tools from APIs with access controls and observability.

> Gram's server release adds an Okta tenant management API alongside new analytics and Explore query services for observability data, admin billing/spend visualization, workload session identity details, and Platform MCP plugin tooling for migrating and importing remote MCP servers; the admin app also lets staff enable Tailscale private access without a rollout flag.

└──▷ WHAT SHIPPED · 10 FEATURES most completely described first ? what's the number?

01 Okta tenant connection management API NEW

85

Adds the `identityProviderConnections` management API for connecting an organization's Okta tenant, gated behind the `okta-connections` flag and requiring `GRAM_IDENTITY_PROVIDER_SIGNING_CREDENTIAL_ID` to be set. Also adds `syncApplications` and `listApplications` on the same API to reconcile each verified Okta connection's applications and assignments into a snapshot on a schedule.

— Names API, flag, env var and reconciliation endpoints. server@2.9.0

02 Structured 401 for invalid workload tokens IMPROVED

85

Revoked, expired, or agent-less workload session tokens on MCP requests now receive `401` with `WWW-Authenticate: Bearer ..., error='invalid_token'` so clients drop and refresh the token instead of retrying.

— Gives exact status code, header value and client behaviour change. server@2.9.0

03 Explore workspace for saved analytics queries NEW

83

Adds the `explore` service for saved queries, with per-save catalog validation, project-write-access enforcement for deleting others' queries, and full audit logging. Adds the Explore page (preview) to Observability navigation behind the `gram-explore` rollout flag, with a query builder generated from the analytics catalog that reconfigures measures, filters, and breakdowns per dataset; filters show available dimension values within the current window with counts while still accepting typed values not in the list; queries run only on explicit Run, scanning the full window with no side-effect evaluations, open on the last 24 hours, place filters above measures, label results panels with the answered query, and render timeseries with a summary table beneath.

— Backend service and UI mechanics both detailed with flag name. dashboard@0.122.0

server@2.9.0

04 **Analytics catalog query service** NEW 80

Adds the `analytics` service: `analytics.query` runs grouped or ungrouped queries against a catalog dataset by field name, `analytics.describe` serves the catalog the query builder is generated from, and `analytics.dimensionValues` returns the most-frequent values a dimension holds inside a window, resolved after the dataset collapses its observations.

— Names three service methods and their mechanism. server@2.9.0

05 **Platform MCP plugin migration and import tooling** NEW 78

Adds the `migrate-mcp-between-projects` skill to the Platform MCP plugin, walking an administrator through redeploying a Platform-managed MCP server from a source project to a target project in the same organization. Adds an optional second step in Platform MCP setup for importing existing remote MCP servers from Claude Code into Speakeasy management, with discovery consent, catalog matching by endpoint or provider/name, and per-server registration reporting.

— Names the skill and setup step mechanics in detail. server@2.9.0

06 **Admin billing and spend visualization** NEW 73

Adds an admin spend breakdown API and organization billing visualization for all account types, including organizations without a Stripe subscription, with product selection, billing-cycle and custom date ranges, daily/weekly/monthly grouping, and cumulative views. Adds a metered-product spend breakdown endpoint and stacked billing chart covering agent session storage, per-scanner risk scanning, and MCP egress, restricted to PAYG organizations (other plans return `unsupported_plan`).

— Names metrics, grouping options and plan-restriction error code. server@2.9.0

07 **Remote MCP probe endpoint** NEW 70

Adds a structured Remote MCP `probeURL` endpoint returning availability, authentication, invalid-response, and unreachable outcomes, while retaining the deprecated `verifyURL` endpoint for compatibility.

— Names endpoint and its outcome states. server@2.9.0

08 **Workload principal type in user sessions** NEW 65

Adds a `workload` principal type: `userSessions.list` now returns a `workload` object for `workload:` subjects carrying the workload issuer's name and URL, the external subject, the assigned agent and its state, and all active admissions.

— Names API method and full set of returned fields. server@2.9.0

THINNER COVERAGE BELOW

09 **Tailscale private access toggle for admins** IMPROVED

50

Enables staff to enable Tailscale private access for an organization directly from the admin feature controls, without requiring a separate PostHog rollout flag.

— Names the mechanism change but description is brief. admin@0.5.0

10 **Admin support coverage matrix** NEW

45

Adds a shared admin support matrix with CSV import, editable coverage, and integration recommendations that include remaining coverage gaps.

— Describes feature in prose without named surfaces. server@2.9.0

server@2.9.0 NOTES ↗ LATEST STABLE

gram server 2.9.0 adds Okta tenant management API, analytics/explore services, workload session labeling, and MCP probe endpoint

└──▷ GET THIS VERSION ▾

```
$ git clone --branch server@2.9.0 https://github.com/speakeasy-api/gram.git
# already have the repo? check out this version:
$ git checkout server@2.9.0
```

Adds the `identityProviderConnections` management API for connecting an organization's Okta tenant, gated behind the `okta-connections` flag and requiring `GRAM_IDENTITY_PROVIDER_SIGNING_CREDENTIAL_ID` to be set.

Adds `syncApplications` and `listApplications` on the `identityProviderConnections` API to reconcile each verified Okta connection's applications and assignments into a snapshot on a schedule.

Adds the `analytics` service: `analytics.query` runs grouped or ungrouped queries against a catalog dataset by field name, and `analytics.describe` serves the catalog the query builder is generated from.

Adds `analytics.dimensionValues` to return the most-frequent values a dimension holds inside a window, resolved after the dataset collapses its observations.

Adds the `explore` service for saved queries, with per-save catalog validation, project-write-access enforcement for deleting others' queries, and full audit logging.

+8 more

dashboard@0.122.0 NOTES ↗ STABLE

gram: Adds the Explore page (preview) to Observability navigation behind the `gram-explore` rollout flag, with a query builder generated from the analytics catalog that reconfigures measures, filters, and breakdowns per dataset.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch dashboard@0.122.0 https://github.com/speakeasy-api/gram.git
# already have the repo? check out this version:
$ git checkout dashboard@0.122.0
```

Explore filters now show available dimension values within the current window with counts, and still accept typed values not in the list.

Explore runs queries only on explicit Run, scanning the full window with no side-effect evaluations; opens on the last 24 hours, places filters above measures, labels results panels with the answered query, and renders timeseries with a summary table beneath.

WAS THIS USEFUL? 👍 👎

Composio

Feedback ⓘ @composio/cli@0.4.2-beta.399 NOTES ↗ RANK

Composio connects AI agents to external tools and services with managed authentication and integrations.

> Composio expanded its toolkit catalog with seven new integrations and shipped several SDK improvements for session management, including a new connection-reconciliation method and a TypeSafe Jev provider.

└──▷ **WHAT SHIPPED · 4 FEATURES** most completely described first ? what's the number?

01 **session.update() returns session config** IMPROVED 65

session.update() now returns the session config object, letting callers inspect the updated state without making a separate fetch call.

– Clear before/after behavior with exact method named. @composio/cli@0.4.2-beta.399

THINNER COVERAGE BELOW

02 **session.ensureConnected() connection reconciliation** NEW 55

Adds session.ensureConnected() to reconcile connection state in the SDK, available in both the TypeScript and Python SDKs.

– Names exact method but mechanism is only briefly described.

@composio/cli@0.4.2-beta.399

03 **Seven new toolkit integrations** NEW 45

Adds seven new toolkit integrations: Amplitude MCP, Clay, Greptile MCP, Inspo MCP, Jev, Modem MCP, and Ramp MCP.

– Names all seven toolkits but gives no usage detail. product docs

04 **TypeSafe Jev provider for SDKs** NEW 35

Adds a TypeSafe Jev provider for both the TypeScript and Python SDKs.

– Just names the addition with no further mechanism given. @composio/cli@0.4.2-beta.399

@composio/cli@0.4.2-beta.399 NOTES ↗ PRE-RELEASE

Composio adds TypeSafe Jev provider, session config return from update(), and session.ensureConnected() for TypeScript and Python

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch @composio/cli@0.4.2-beta.399
https://github.com/ComposioHQ/composio.git
# already have the repo? check out this version:
$ git checkout @composio/cli@0.4.2-beta.399
```

Adds session.ensureConnected() to reconcile connection state in the SDK (TypeScript and Python).

Makes session.update() return the session config object, enabling callers to inspect updated state without a separate fetch.

Adds TypeSafe Jev provider for both TypeScript and Python SDKs.

WAS THIS USEFUL?  

DEPLOY

◆ Local LLM Runtimes

Comfy Org ComfyUI

Feedback ⓘ v0.37.0 NOTES ↗ RANK

ComfyUI is a node-graph interface, API, and backend for running diffusion models, chaining generation steps into reusable visual workflows.

> ComfyUI expanded its node library with FastH3 video, YuE2 music, Marigold V2, Tripo 3D, Pruna video, and Bria/Flux video-editing nodes, added generic graph loops and automatic `--fast-disk` detection, and updated its Gemini, Meshy, and OpenAI partner nodes alongside Qwen inference acceleration.

└──▷ WHAT SHIPPED · 17 FEATURES most completely described first ? what's the number?

01 **Automatic fast-disk detection** IMPROVED 65

ComfyUI now auto-detects fast disks and enables the `--fast-disk` flag automatically when the disk qualifies, removing the need for manual configuration; an option remains to disable it for debugging.

— Names the flag and before/after behaviour, but not the disable syntax. v0.37.0

02 **Generic Loops in the graph** NEW 65

Adds Generic Loops support to loop over lists in the graph, carrying heterogeneous items through every iteration, with loop boundaries declared in node schemas.

— Explains mechanism but gives no example workflow. product docs

03 **Flux Video Edit node** NEW 65

Adds a Flux Video Edit node to edit a source clip from a text instruction: remove or replace objects, restyle footage, and change on-screen text or dialogue.

— Lists concrete edit capabilities but no parameter list. product docs

04 **FastVideo FastH3 video nodes** NEW 60

Adds FastVideo `FastH3` nodes for 8-step distilled MiniMax H3 text-to-video and image-to-video generation with native audio.

— Names node and specs but no usage path. product docs

05 **YuE2 music generation with CFG control** IMPROVED 60

Adds YuE2 nodes for music generation from style prompts and lyrics, including a Generate Music node; a subsequent release adds CFG control to the `YuE2 Generate Music` node for tighter guidance over output.

— Names node and added control, no parameter range given. v0.37.0

06 **Pruna P-Video-2 video nodes** NEW 60

Adds Pruna P-Video-2 nodes for text-to-video and image-to-video generation at 720p/1080p with a draft mode.

— Names resolutions and draft mode, no usage steps. product docs

07 **Templates v0.11.62 refresh** IMPROVED 60

Templates v0.11.62 adds FastH3 video, Marigold V2 estimation, and YuE2 music templates, and updates the Tripo, Gemini, and Ideogram workflow templates.

— Names template set and version, not exact file names. product docs

08 **Partner node updates for Gemini, Meshy, and OpenAI** NEW 60

Adds `GeminiNodeV3` node for Gemini V3 integration in the Gemini Partner Node (the V2 node is now deprecated); adds Meshy 7.1 support in the Meshy Partner Node; adds transparent background support for GPT Image 2 in the OpenAI Partner Node.

— Names each partner node update but not the mechanism per node. v0.37.0

09 **CUDA graphs and w4a8 GEMV for Qwen models** IMPROVED 60

Adds CUDA graphs and w4a8 GEMV support for Qwen 3/3.5/3.8 models, accelerating inference.

— Names technique and models but gives no benchmark numbers. v0.37.0

THINNER COVERAGE BELOW

10 **Linear color space option** NEW 50

Adds a Linear color space option to the `ImageColorSpace` node for linear color space processing.

— Names node and option but no usage detail. product docs

11 **Marigold V2 estimation nodes** NEW 50

Adds Marigold V2 nodes for depth, surface normal, and albedo estimation from a single input image.

— Describes outputs but no workflow steps. product docs

12 **Tripo P2 3D generation nodes** NEW 50

Adds Tripo P2 nodes for text-to-3D, image-to-3D, and multiview-to-3D model generation.

— Names modalities but no workflow steps. product docs

13 **AMD Windows VA quota bump** IMPROVED 45

Bumps the AMD Windows VA quota to 4TB to support larger workloads.

— Gives the new limit but no configuration steps. product docs

14 **Text encoder GPU placement in dynamic VRAM mode** IMPROVED 45

Always places the text encoder on GPU when dynamic VRAM mode is active, improving throughput.

— Describes the behaviour change but no config flag named. v0.37.0

15 **Video Concatenate node** NEW 40

Adds a Video Concatenate node to join multiple video clips end to end.

— Thin description of a single node. product docs

16 **Bria edit suite** NEW 40

Adds a Bria edit suite with new image edit nodes and a Video Eraser node.

— Names suite and one node, rest left vague. product docs

17 **Partner job progress estimates** NEW 40

Adds partner progress estimates that show remaining time based on server duration estimates while partner jobs are queued and running.

— Explains behaviour but no named UI location. product docs

v0.37.0 NOTES ↗ LATEST STABLE

ComfyUI v0.37.0 adds MoGe 3, Qwen-Image 2.1, CFG control for YuE2, transparent background for GPT Image 2, and Meshy 7.1 support.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.37.0 https://github.com/Comfy-Org/ComfyUI.git
# already have the repo? check out this version:
$ git checkout v0.37.0
```

Adds `--fast-disk` auto-detection: ComfyUI now detects fast disks and enables the flag automatically, with an option to disable it for debugging.

Adds transparent background support for GPT Image 2 in the OpenAI Partner Node.

Adds CFG control to the YuE2 Generate Music node.

> Adds Meshy 7.1 support in the Meshy Partner Node.
>
> Adds CUDA graphs and w4a8 GEMV support for Qwen 3/3.5/3.8 models, accelerating inference.
>
> +1 more

WAS THIS USEFUL?  

◆ Agentic Infrastructure

HeyGen HyperFrames

Feedback ⓘ v0.8.58 NOTES ↗ RANK

HyperFrames is an open-source HTML-to-video renderer that runs in AI-agent workflows.

> HyperFrames added CLI-driven timeline editing and expanded Studio's host-app embedding surface with playback-handle export, read-only locking and CSS-token theming, alongside several smaller Studio debugging and layout aids.

└──▷ **WHAT SHIPPED · 5 FEATURES** most completely described first ? what's the number?

01 **Host app embedding controls in Studio** NEW 65

Studio now exports its player handle so host applications can programmatically control playback, can be locked to a read-only preview mode (removing the need for a separate edit guard in the embedding application), and supports theming of timeline colours through CSS tokens (eliminating the need for custom colour overrides in host apps).

– Names player handle, read-only mode and CSS tokens with purpose v0.8.56 v0.8.58

THINNER COVERAGE BELOW

02 **Per-page tab ID and hash-validation verdict** NEW 50

Adds a per-page tab ID and exposes the hash-validation verdict in Studio for tracking and debugging page identity.

– Names the tab ID and verdict but not where they surface v0.8.57

03 **CLI timeline editing verbs** NEW 35

Adds timeline editing verbs to the CLI, letting users manipulate the timeline directly instead of through Studio's UI.

– Names the CLI surface but no exact verb syntax given v0.8.58

04 **Compose preview overlay guides in Studio** NEW 35

Adds compose preview overlay guides in Studio to aid layout and composition alignment.

– States purpose but no mechanism or UI path detail v0.8.57

05 **Surfaced project info on failed reads** IMPROVED 35

Studio now surfaces which project it is actually serving when a read 404s, instead of failing silently.

– Describes before/after but no specific surface named v0.8.57

v0.8.58 NOTES ↗ LATEST STABLE

HyperFrames v0.8.58 adds CLI timeline editing verbs and a Studio player handle export for host apps.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.8.58 https://github.com/heygen-com/hyperframes.git
# already have the repo? check out this version:
$ git checkout v0.8.58
```

Adds timeline editing verbs to the CLI for direct timeline manipulation.

Studio exports its player handle so host applications can programmatically control playback.

v0.8.57 NOTES ↗ STABLE

HyperFrames v0.8.57 adds compose preview overlay guides, per-page tab IDs with hash-validation verdicts, and a speed-ramp rate lane with presets for clip editing.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.8.57 https://github.com/heygen-com/hyperframes.git
# already have the repo? check out this version:
$ git checkout v0.8.57
```

Adds a per-page tab ID and exposes the hash-validation verdict in Studio for tracking and debugging page identity.

Adds compose preview overlay guides in Studio to aid layout and composition alignment.

Surfaces the project Studio is actually serving when a read 404s, instead of failing silently.

v0.8.56 NOTES ↗ STABLE

HyperFrames v0.8.56 lets host apps lock Studio previews to read-only and restyle the timeline via CSS tokens.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.8.56 https://github.com/heygen-com/hyperframes.git
# already have the repo? check out this version:
$ git checkout v0.8.56
```

Enables host apps to make the Studio preview read-only, removing the need for a separate edit guard in the embedding application.

Supports theming of Studio timeline colours through CSS tokens, eliminating the need for custom colour overrides in host apps.

WAS THIS USEFUL? 👍 👎

Snowflake Cortex AI

Feedback ⓘ

Sep 21, 2026: Cortex Agents Compact API (Preview) NOTES ↗ RANK

Snowflake Cortex AI provides managed language-model functions, search, and agents that work with data governed in Snowflake.

> Snowflake Cortex AI added a new `agent:compact` endpoint to Cortex Agents, letting long conversations be summarized to save tokens and stay within model context limits.

└──▷ **WHAT SHIPPED · 1 FEATURE** most completely described first ? what's the number?

01 **Conversation compaction endpoint for Cortex Agents** NEW

80

Adds the `agent:compact` endpoint to summarize a conversation and pass the compact representation to subsequent `agent:run` requests, reducing token consumption and keeping conversations within the model context window. It supports both inline transcripts and conversations stored in Cortex Agent threads as input to `agent:compact`.

— Names endpoint and mechanism, no usage example or limits given

Sep 21, 2026: Cortex Agents Compact API…

> Sep 21, 2026: Cortex Agents Compact API (Preview) NOTES ↗ LATEST STABLE
>
> Cortex Agents gains an `agent:compact` endpoint to summarize conversation history and reduce token consumption.
>
> Adds the `agent:compact` endpoint to summarize a conversation and pass the compact representation to subsequent `agent:run` requests, reducing token consumption and keeping conversations within the model context window.
>
> Supports both inline transcripts and conversations stored in Cortex Agent threads as input to `agent:compact`.

WAS THIS USEFUL? 👍 👎

Eigen Labs Darkbloom

Feedback ⓘ v0.9.7 NOTES ↗ CODE ↗

RANK

Private Inference Network on Idle Macs

> Darkbloom v0.9.7 lets providers join its idle-Mac inference network via App Attest without MDM enrollment, adds per-account first-content SLA configuration, kicks off a Bonsai campaign with claimable model token grants, and speeds up full-integrity model verification by default.

└──▷ **WHAT SHIPPED · 4 FEATURES** most completely described first ? what's the number?

01 First-content SLA configuration on coordinator NEW

90

Adds the `EIGENINFERENCE_MODEL_FIRST_CONTENT_SLAS` environment variable to configure per-model first-content SLA timing on the coordinator, and `EIGENINFERENCE_FIRST_CONTENT_SLA_ACCOUNTS` to scope that SLA enforcement to specific authenticated accounts so other direct users and service accounts have no first-content timeout.

Restrict first-content SLA enforcement to a specific account so other users and service accounts experience no first-content timeout.

```
$ export EIGENINFERENCE_FIRST_CONTENT_SLA_ACCOUNTS=<account-id>
export EIGENINFERENCE_MODEL_FIRST_CONTENT_SLAS='{"bonsai-2-27b":
{"first_content_ms":9000}}'
./coordinator
```

— Names both env vars and gives a runnable example. v0.9.7

02 Bonsai campaign token grants and model promotion NEW

65

Enables claimable model token grants for the Bonsai campaign — 150M tokens each for the first 250 eligible accounts — redeemable through the console, qualifies Bonsai 2 27B for lossless performance and API stability, and prioritizes Bonsai models first in the chat model selector UI with a Free badge shown after tokens are claimed.

— Names concrete grant size, model, and console redemption path. v0.9.7

THINNER COVERAGE BELOW

03 Faster full-integrity model verification IMPROVED

55

Enables faster full-integrity model verification by default using reusable-buffer reads and up to four independent file readers.

— Mechanism named but no config exposed to act on. v0.9.7

04 **App Attest enrollment without Darkbloom MDM** NEW 40

Allows qualified providers to join the network through App Attest without enrolling in Darkbloom MDM, including Macs that are already under employer management.

— Describes the change but no enrollment steps given. v0.9.7

v0.9.7 NOTES ↗ CODE ↗ LATEST STABLE

Darkbloom v0.9.7 adds MDM-optional App Attest providers, account-scoped first-content SLAs, claimable model token grants, and faster model verification by default.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.9.7 https://github.com/Layr-Labs/d-inference.git
# already have the repo? check out this version:
$ git checkout v0.9.7
```

└──▷ TRY IT

Restrict first-content SLA enforcement to a specific account so other users and service accounts experience no first-content timeout.

```
$ export EIGENINFERENCE_FIRST_CONTENT_SLA_ACCOUNTS=<account-id>
export EIGENINFERENCE_MODEL_FIRST_CONTENT_SLAS='{"bonsai-2-27b":{"first_content_ms":9000}}'
./coordinator
```

Adds `EIGENINFERENCE_MODEL_FIRST_CONTENT_SLAS` environment variable to configure per-model first-content SLA timing on the coordinator.

Scopes first-content SLAs to specific authenticated accounts via `EIGENINFERENCE_FIRST_CONTENT_SLA_ACCOUNTS`, so direct users and other service accounts have no first-content timeout.

Allows qualified providers to join the network through App Attest without enrolling in Darkbloom MDM, including Macs under employer management.

Enables claimable model token grants (Bonsai campaign: 150M tokens each for the first 250 eligible accounts) redeemable through the console.

> Enables faster full-integrity model verification using reusable-buffer reads and up to four independent file readers by default.
>
> +2 more

WAS THIS USEFUL? 👍 👎

Nebius AI Cloud

Feedback ⓘ September 14–20, 2026 (2026-09-21)

BLOG ↗ RANK

Nebius AI Cloud provides GPU compute, managed Kubernetes, distributed training infrastructure, and serverless AI workloads.

> Nebius AI Cloud shipped a full Audit Logs system spanning CLI and console, alongside networking and platform additions including DNS zones for virtual networks, a new Madrid region, multi-cloud SkyPilot bursting, and a serverless Devlabs deployment type.

└──▷ **WHAT SHIPPED · 6 FEATURES** most completely described first ? what's the number?

01 **Audit Logs with CLI, console and metrics** NEW 95

Adds the `nebius audit v2 audit-event list` CLI command to retrieve audit events for a tenant, supporting `--parent-id`, `--start`, `--end`, `--region`, and `--event-type` flags with ISO 8601 time bounds, including `--event-type data_plane` to pull data-plane events (e.g. Object Storage reads/writes) separately from default control-plane events. Adds an Audit Logs section under Administration → Audit Logs in the web console with a Logs tab for viewing/filtering control-plane and data-plane event JSONs by region and time, and a Metrics tab showing events processed per second by region/time plus quota usage for data-plane events. Events are represented as JSON conforming to the CloudEvents specification, covering both control-plane and data-plane operations.

Pull all audit events for a tenant during a specific month to investigate activity across a region.

```
$ nebius audit v2 audit-event list \
  --parent-id <tenant_ID> \
  --start 2025-03-01T00:00:00Z \
  --end 2025-04-01T00:00:00Z \
  --region <region_ID>
```

Retrieve data-plane events (e.g. Object Storage reads/writes) for a time window when investigating unauthorized data access.

```
$ nebius audit v2 audit-event list \
  --parent-id <tenant_ID> \
  --start 2025-03-01T00:00:00Z \
  --end 2025-04-01T00:00:00Z \
  --event-type data_plane \
  --region <region_ID>
```

View and filter audit event JSONs in the console when you need a quick visual triage of recent control-plane or data-plane activity.

📍 In the web console, go to Administration → Audit Logs. On the Logs tab, select the event type (Control plane or Data plane), select the region, click Filters to narrow by time period or other criteria, then click the expand icon on any event row to view its full JSON.

— Exact CLI command, flags and console path all given product docs

02 **Multi-cloud SkyPilot cluster connections** NEW 70

Enables connecting a SkyPilot cluster to third-party clouds — Google Cloud, AWS, or Runpod — by providing credentials at cluster create or edit time, allowing workloads to run across multiple clouds.

Run GPU workloads that burst across Nebius, AWS, and Google Cloud from a single SkyPilot cluster definition.

📍 In the Nebius AI Cloud console, go to SkyPilot › your cluster › Edit, enter your AWS and Google Cloud credentials, then submit jobs — SkyPilot schedules them across all connected clouds.

— Names three providers and console steps, no API endpoint

September 14–20, 2026 (2026-09-21)

03 **DNS zones for virtual networks** NEW 65

Adds DNS in Virtual Networks: create DNS zones attached to a network and add records that resolve only from workloads in that network, giving VMs stable internal endpoints under a custom domain name.

Give VMs in a private network stable internal DNS names so services can reach each other by hostname instead of IP address.

📍 In the Nebius AI Cloud console, go to Virtual Networks › your network › DNS Zones, click Create DNS zone , enter your chosen domain name, then add A or CNAME records pointing to your VM internal IPs.

– Console navigation given, no CLI or API name September 14–20, 2026 (2026-09-21)

04 **Serverless AI Devlabs deployment type** NEW 65

Adds Serverless AI Devlabs as a new deployment type (alongside jobs and endpoints) that runs container images as interactive, browser-based development environments for coding, prototyping, and debugging.

Spin up an interactive browser-based coding environment from a container image without managing a persistent endpoint or job queue.

> 📍 In the Nebius AI Cloud console, go to Serverless › Devlabs, click Create Devlab, select your container image, and launch — the environment opens in the browser for interactive development.

– Console path given, no config or CLI named September 14–20, 2026 (2026-09-21)

THINNER COVERAGE BELOW

05 **New Madrid private region** NEW 30

Adds the new `eu-south1` (Spain, Madrid) private region.

– Bare region name, no usage detail September 14–20, 2026 (2026-09-21)

06 **Increased shared filesystem specification limits** IMPROVED 15

Increases shared filesystem specification values.

– No concrete numbers or limits given September 14–20, 2026 (2026-09-21)

September 14–20, 2026 (2026-09-21) BLOG ↗ LATEST STABLE

Nebius AI AI Cloud adds DNS for virtual networks, Madrid region, multi-cloud SkyPilot, and serverless Devlabs

└──▷ **HOW TO FIND IT**

Give VMs in a private network stable internal DNS names so services can reach each other by hostname instead of IP address.

> 📍 In the Nebius AI Cloud console, go to Virtual Networks › your network › DNS Zones, click Create DNS zone, enter your chosen domain name, then add A or CNAME records pointing to your VM internal IPs.

Spin up an interactive browser-based coding environment from a container image without managing a persistent endpoint or job queue.

> 📍 In the Nebius AI Cloud console, go to Serverless › Devlabs, click Create Devlab, select your container image, and launch — the environment opens in the browser for

interactive development.

Run GPU workloads that burst across Nebius, AWS, and Google Cloud from a single SkyPilot cluster definition.

📍 In the Nebius AI Cloud console, go to SkyPilot › your cluster › Edit, enter your AWS and Google Cloud credentials, then submit jobs — SkyPilot schedules them across all connected clouds.

Adds DNS in Virtual Networks: create DNS zones attached to a network and add records that resolve only from workloads in that network, giving VMs stable internal endpoints under a custom domain name.

Adds the new `eu-south1` (Spain, Madrid) private region.

Enables connecting a SkyPilot cluster to third-party clouds — Google Cloud, AWS, or Runpod — by providing credentials at cluster create or edit time, allowing workloads to run across multiple clouds.

Adds Serverless AI Devlabs as a new deployment type (alongside jobs and endpoints) that runs container images as interactive, browser-based development environments for coding, prototyping, and debugging.

Increases shared filesystem specification values.

WAS THIS USEFUL? 👍 👎

◆ Local LLM Hardware

Minisforum MS-S1 Max Feedback ⓘ

back in stock · Sep 20, 2026 STORE ↗

A Ryzen AI Max+ 395 mini workstation with 80 Gbps USB4 v2 and 2U rack-mount support, aimed at multi-unit local AI clusters.

MS-S1 Max · 128 GB (ms-s1-max-128) back in stock at `store.minisforum.com` on 20 Sep 2026.

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

Feedback ⓘ snapshot-20260921 NOTES ↗

LangSmith provides tracing, evaluation, and deployment tools for LLM applications.

└──▷ TRY IT

Pull thread statistics including latency, token usage, and cost for a tracing project to feed into a security posture dashboard or cost audit.

```
$ curl -X GET 'https://api.smith.langchain.com/v2/threads/stats?
project_id=<project_id>' \
  -H 'x-api-key: <LANGSMITH_API_KEY>'
```

Give self-hosted operators earlier notice of an approaching license expiration to avoid unexpected access interruptions.

.env

```
LICENSE_EXPIRATION_WARNING_DAYS=14
```

Tag a custom OpenAI-compatible model with its provider so the LLM Gateway records accurate trace metadata and applies provider-specific pricing.

json

```
{
  "model": "my-custom-model",
  "ls_provider": "openai",
  "base_url": "https://my-gateway.example.com/v1"
}
```

└──▷ BREAKING ON UPGRADE

! Dataset CSV and JSONL exports now require both dataset read and download permissions; API callers that previously exported with read-only tokens will receive a forbidden response.

WAS THIS USEFUL? 👍 👎

Braintrust

Feedback ⓘ announced · Sep 21, 2026 BLOG ↗ RANK

Braintrust provides evaluation, tracing, and improvement workflows for AI applications.

> Braintrust's headline addition is a native Harbor plugin that syncs sandboxed agent eval runs, traces, and scores directly into Braintrust projects, alongside a batch of Loop-powered features (Patterns, scheduled automations, thread sharing), new coding-session tracing for Grok and Google Antigravity, and a faster streaming log search engine.

└──▷ **WHAT SHIPPED · 5 FEATURES** most completely described first ? what's the number?

01 **Harbor agent eval sync via Braintrust plugin** NEW 95

Adds `--plugin braintrust` and `--plugin-kwarg project_name=<name>` flags to `harbor run` for syncing sandboxed agent eval jobs directly to a Braintrust project, with `HARBOR_BRAINTRUST_PROJECT` and `HARBOR_BRAINTRUST_*` environment variables as alternatives to `--plugin-kwarg` arguments. The `braintrust.integrations.harbor.backfill_job` Python function syncs a previously completed Harbor job directory without re-running the agent or verifier, writing a `braintrust-sync.json` manifest into each job directory recording what was synced. Agent trajectories in Harbor's ATIF format (tool calls and model turns) are synced as traces nested under a single eval span alongside scores, and each verifier reward becomes its own chartable, filterable Braintrust score field.

Run a Harbor agent eval and sync results live to a Braintrust project for cross-run comparison and sharing.

```
$ uv run harbor run \
  --path task \
  --agent terminus-2 \
  --model openai/gpt-4.1-mini \
  --job-name braintrust-harbor-example \
  --plugin braintrust \
  --plugin-kwarg project_name=example-harbor
```

Backfill an already-completed Harbor job into Braintrust without re-running the agent or verifier — safe to rerun; reconciles existing rows instead of adding duplicates.

```python
import asyncio
from pathlib import Path
from braintrust.integrations.harbor import backfill_job

asyncio.run(backfill_job(Path("jobs/my-eval"), project_name="my-project"))
```

Set the Braintrust project via environment variable to avoid repeating `--plugin-kwarg` on every `harbor run` invocation.

```
$ export HARBOR_BRAINTRUST_PROJECT=example-harbor

uv run harbor run \
  --path task \
  --agent terminus-2 \
  --model openai/gpt-4.1-mini \
  --job-name braintrust-harbor-example \
  --plugin braintrust
```

— Full flag/function/env-var surface with runnable examples `launch-20260921-4ccbd7e5`

THINNER COVERAGE BELOW

02 **Scheduled Loop automations with Slack/webhook output** NEW

55

Adds Loop automations — scheduled runs with their own instruction, model, and write permissions — that leave a read-only thread and can post results to a Slack channel or webhook. Requires data plane v2.12.0+.

— Named mechanism but no configuration steps given `snapshot-20260921`

03 **Blind human reviews project setting** NEW

55

Adds the Blind human reviews project setting to hide peer scores, comments, and aggregates from a reviewer until they submit their own scores.

— Names exact setting but no navigation detail `snapshot-20260921`

04 **Patterns automated failure-mode detection** NEW 50

Adds Patterns, a scheduled Loop-powered feature that automatically investigates traces for recurring failure modes and trends, saves findings with supporting traces and suggested fixes, and lets you create a scorer or classifier directly from a finding. Requires data plane v2.13+.

— Mechanism described but no setup path or example snapshot-20260921

05 **Loop thread sharing for collaboration** NEW 35

Adds Loop thread sharing, allowing project members to collaborate on an investigation with view-only or chat access. Requires data plane v2.15.0+.

— Brief description, no usage path given snapshot-20260921

└──▷ **BREAKING ON UPGRADE**

! On Braintrust-hosted deployments, project providers now override organization providers only for the models they explicitly serve; organization providers can serve other models with their own credentials. If you relied on project providers to restrict model access or spending across all models, review organization provider model selections and permissions.

announced · Sep 21, 2026 BLOG ↗

Braintrust gains a native Harbor plugin to sync sandboxed agent eval results — runs, traces, and scores — into Braintrust projects.

└──▷ **TRY IT**

Run a Harbor agent eval and sync results live to a Braintrust project for cross-run comparison and sharing.

```
$ uv run harbor run \
  --path task \
  --agent terminus-2 \
  --model openai/gpt-4.1-mini \
  --job-name braintrust-harbor-example \
  --plugin braintrust \
  --plugin-kwarg project_name=example-harbor
```

Backfill an already-completed Harbor job into Braintrust without re-running the agent or verifier — safe to rerun; reconciles existing rows instead of adding duplicates.

```python
import asyncio
from pathlib import Path
from braintrust.integrations.harbor import backfill_job

asyncio.run(backfill_job(Path("jobs/my-eval"), project_name="my-project"))
```

Set the Braintrust project via environment variable to avoid repeating `--plugin-kwarg` on every `harbor run` invocation.

```
$ export HARBOR_BRAINTRUST_PROJECT=example-harbor

uv run harbor run \
  --path task \
  --agent terminus-2 \
  --model openai/gpt-4.1-mini \
  --job-name braintrust-harbor-example \
  --plugin braintrust
```

Adds `--plugin braintrust` and `--plugin-kwarg project_name=<name>` flags to `harbor run` for syncing sandboxed agent eval jobs directly to a Braintrust project.

Supports `HARBOR_BRAINTRUST_PROJECT` and `HARBOR_BRAINTRUST_*` environment variables as alternatives to `--plugin-kwarg` arguments on `harbor run`.

Adds `braintrust.integrations.harbor.backfill_job` Python function to sync a previously completed Harbor job directory to Braintrust without re-running the agent or verifier.

Writes a `braintrust-sync.json` manifest into each Harbor job directory recording what was synced to Braintrust.

Syncs agent trajectories in Harbor's ATIF format — tool calls and model turns — as traces nested under a single eval span alongside the scores they produced.

+1 more

snapshot-20260921 NOTES ↗ LATEST STABLE

Braintrust: Adds Patterns, a scheduled Loop-powered feature that automatically investigates traces for recurring failure modes and trends, saves findings with supporting traces and suggested fixes, and lets you create a scorer or classifier directly from a finding; requires data plane v2.13+.

└──▷ HOW TO FIND IT

Enable the columnar metric store on a high-volume project so Logs page queries read far less data when filtering and sorting traces.

📍 In the project, go to Settings › Log search optimization and enable `Metric columnstore`.

Adds Loop automations — scheduled runs with their own instruction, model, and write permissions — that leave a read-only thread and can post results to a Slack channel or webhook; requires data plane v2.12.0+.

Adds the `Blind human reviews` project setting to hide peer scores, comments, and aggregates from a reviewer until they submit their own scores.

Adds Loop thread sharing, allowing project members to collaborate on an investigation with view-only or chat access; requires data plane v2.15.0+.

Adds Grok and Google Antigravity session tracing via the `bt` CLI v0.19.3+, capturing each coding session as a Braintrust trace without interrupting the agent if tracing fails.

Brainstore log search now streams results from an asynchronous execution engine as segments finish scanning, covering full-text, phrase, and multi-clause ANY_SPAN() trace searches with no configuration changes required; requires data plane v2.13.0+.

└──▷ **BREAKING ON UPGRADE**

! On Braintrust-hosted deployments, project providers now override organization providers only for the models they explicitly serve; organization providers can serve other models with their own credentials. If you relied on project providers to restrict model access or spending across all models, review organization provider model selections and permissions.

WAS THIS USEFUL? 👍 👎

Langfuse

Feedback ⓘ v4.40.0 NOTES ↗ RANK

Langfuse provides tracing, evaluation, and monitoring for LLM applications.

> Langfuse's latest releases link transcript spans to their source traces in the trace-batching pipeline and add several smaller improvements spanning automations, in-app AI backends, the conversation API, and AI gateway tracing.

└──▷ **WHAT SHIPPED · 5 FEATURES** most completely described first ? what's the number?

01 **AI gateway trace visibility and metadata** IMPROVED 45

Adds phase spans to AI gateway request traces to eliminate gaps in request visibility, and records Codex client metadata as agent metadata in the AI gateway.

– Names both additions but no flags or endpoints to act on v4.39.0

02 **Transcript-to-trace linking in trace batching** NEW 40

Links transcript spans to their source traces in the trace-batching pipeline, enabling end-to-end traceability across batched LLM interactions.

– Names the pipeline and mechanism but no config or command given v4.40.0

03 **Conversation history per thread via API** IMPROVED 40

Returns conversation history and current turn data per thread via the API.

– Names the API surface but omits endpoint path or parameters v4.39.0

04 **Vertex AI backend for in-app AI features** NEW 35

Supports Vertex AI as a backend for instance-level in-app AI features.

– Names Vertex AI but no setup steps or config key v4.39.0

05 **Label-based filtering for prompt events in automations** NEW 30

Adds label-based filtering for prompt events in automations, enabling finer-grained trigger control.

– States the capability without detailing labels or UI/config path v4.39.0

v4.40.0 NOTES ↗ LATEST STABLE

Langfuse v4.40.0 links transcript spans to their source traces via trace batching.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch v4.40.0
https://github.com/langfuse/langfuse.git
# already have the repo? check out this version:
$ git checkout v4.40.0
```

Links transcript spans to source traces in the trace-batching pipeline, enabling end-to-end traceability across batched LLM interactions.

v4.39.0 NOTES ↗ STABLE

Langfuse v4.39.0 adds prompt-event label filtering for automations, Vertex AI support for in-app AI features, and conversation history per thread.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v4.39.0
https://github.com/langfuse/langfuse.git
# already have the repo? check out this version:
$ git checkout v4.39.0
```

Adds label-based filtering for prompt events in automations, enabling finer-grained trigger control.

Supports Vertex AI as a backend for instance-level in-app AI features.

Returns conversation history and current turn data per thread via the API.

Adds phase spans to AI gateway request traces to eliminate gaps in request visibility.

Records Codex client metadata as agent metadata in the AI gateway.

WAS THIS USEFUL?  

GOVERN

◆ AI/LLM Security

ai-safe2-framework

Feedback ⓘ 2026_09_20_CLI_0.7.0

NOTES ↗ RANK

Ai-Safe2-Framework provides governance, risk, and compliance controls for securing agentic AI systems and non-human identities.

> AI SAFE² CLI 0.7.0 introduces a multi-harness evidence correlation command with human decision cards, a companion change-monitoring command for agent configuration, and new operational-truth and change-monitor contracts, alongside hardened evidence collection.

└──▷ **WHAT SHIPPED · 6 FEATURES** most completely described first ? what's the number?

01 **Evidence truth correlation command and operational truth contracts**

NEW 93

Adds `safe2 evidence truth` command to correlate harness evidence, revision-bound task receipts, coverage, usage declarations, and completion claims across multiple agent harnesses. Introduces the `operational-truth-source-v1` contract to declare the task, revision, required evidence, receipt bindings, decision owner, assumptions, and exclusions, and the `operational-truth-manifest-v1` contract to record correlated results, decision gate (`hold`, `review`, `ready_for_human_decision`), conflicts, gaps, and evidence boundaries. Also adds a human decision card output via `--card` that translates canonical evidence into a readable summary alongside the canonical JSON output.

— Command, two contracts, and flag all named verbatim. 2026_09_20_CLI_0.7.0

02 **Agent change monitoring command and recognized configuration surfaces**

NEW 88

Adds `safe2 evidence changes` command to run a one-shot inventory of skills and recognized agent configuration locally or in CI, with `--baseline`, `--output`, and `--strict` flags, backed by the `change-monitor-v1` contract which records bounded agent-input inventory and change results without exporting configuration contents. Recognizes agent configuration beneath `.codex`, `.claude`, `.hermes`, and `.openclaw` directories, as well as MCP and OpenClaw configuration and skill packages rooted by `SKILL.md`.

— Command, flags, contract, and recognized paths named. 2026_09_20_CLI_0.7.0

03 **Strict mode for evidence commands** NEW 70

Adds `--strict` mode to `safe2 evidence` commands, which writes evidence artifacts before returning a nonzero status on `review` or `hold` outcomes.

— Flag and exit-status behavior named, no example run shown. 2026_09_20_CLI_0.7.0

04 **Bounded traversal protections in evidence collection** NEW

68

Adds bounded traversal protections with a default limit of 10,000 entries and 1 MB per tracked file, plus symbolic-link, reparse-point, special-file, and file-identity protections.

— Concrete limits given but no configuration path documented. 2026_09_20_CLI_0.7.0

THINNER COVERAGE BELOW

05 **Complete skill-package hashing and evidence publication** NEW

53

Adds complete skill-package hashing including companion files, exclusive complete-before-visible evidence publication, and no-overwrite behavior with identity-aware cleanup.

— Mechanism described but no usage entry point given. 2026_09_20_CLI_0.7.0

06 **Expanded Python version support** IMPROVED

25

Supports Python 3.11, 3.12, 3.13, and 3.14.

— Bare version list with no further detail. 2026_09_20_CLI_0.7.0

2026_09_20_CLI_0.7.0 NOTES ↗ LATEST STABLE

AI SAFE² CLI 0.7.0 adds multi-harness evidence correlation, agent change monitoring, and human decision cards for AI operational truth.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch 2026_09_20_CLI_0.7.0
https://github.com/CyberStrategyInstitute/ai-safe2-framework.git
# already have the repo? check out this version:
$ git checkout 2026_09_20_CLI_0.7.0
```

Adds `safe2 evidence truth` command to correlate harness evidence, revision-bound task receipts, coverage, usage declarations, and completion claims across multiple agent harnesses.

Adds `safe2 evidence changes` command to run a one-shot inventory of skills and recognized agent configuration locally or in CI, with `--baseline`, `--output`, and `--strict` flags.

> Introduces the `operational-truth-source-v1` contract to declare the task, revision, required evidence, receipt bindings, decision owner, assumptions, and exclusions.
>
> Introduces the `operational-truth-manifest-v1` contract to record correlated results, decision gate (`hold`, `review`, `ready_for_human_decision`), conflicts, gaps, and evidence boundaries.
>
> Introduces the `change-monitor-v1` contract to record bounded agent-input inventory and change results without exporting configuration contents.
>
> +6 more

WAS THIS USEFUL?  

agent-safe-pipeline

Feedback ⓘ v0.3.2 NOTES ↗ RANK

Agent Safe Pipeline provides a security-focused framework for building AI agent workflows with policy checks and safety controls.

> Agent Safe Pipeline's latest releases add hosted shadow-mode enforcement testing, one-command workspace provisioning, a GitHub Actions gate action, and new destination governance and scheduled shadow reporting.

└──▷ **WHAT SHIPPED · 5 FEATURES** most completely described first ? what's the number?

01 Hosted shadow-mode testing via agentsafe test --hosted NEW

85

`agentsafe test --hosted` replays requests against a Decionis-hosted synthetic target in shadow mode, producing one signed Decision Dossier per consequential request, so a team can see which actions would have been held or refused before turning enforcement on. It runs as a boundary test with Decionis deciding and requires no local policy configuration, putting a workspace's first governed action one command away.

Run shadow-mode enforcement testing against Decionis-hosted infrastructure to see which actions would have been held or refused before turning enforcement on.

```
$ agentsafe login --provision && agentsafe test --hosted
```

Validate what hosted enforcement would have held or refused before flipping enforcement on — no account needed.

```
$ agentsafe login --provision && agentsafe test --hosted
```

Bootstrap a workspace's first governed action — run the hosted boundary test without any local policy configuration.

```
$ agentsafe test --hosted
```

— Names exact command and mechanism (Decision Dossier, boundary test). v0.3.2

02 Workspace provisioning via agentsafe login --provision NEW

70

`agentsafe login --provision` mints a free Decionis workspace in a single command with no account required, and provisions credentials as part of the login flow for a fresh environment with no prior session.

Run shadow-mode enforcement testing against Decionis-hosted infrastructure to see which actions would have been held or refused before turning enforcement on.

```
$ agentsafe login --provision && agentsafe test --hosted
```

Validate what hosted enforcement would have held or refused before flipping enforcement on — no account needed.

```
$ agentsafe login --provision && agentsafe test --hosted
```

Provision credentials during login in a fresh environment where no prior session exists.

```
$ agentsafe login --provision
```

— Exact command given but mechanism only lightly described. v0.3.2

03 **CI gate action decionis/govern@v1** NEW 70

The `decionis/govern@v1` GitHub Actions step, added via `- uses: decionis/govern@v1` in a workflow file such as `.github/workflows/your-workflow.yml`, gates outbound agent actions in a CI pipeline so every action is governed by the Decionis authority the workload trusts.

Gate outbound agent actions in a CI pipeline by adding the Decionis govern action to a GitHub Actions workflow.

```
$ - uses: decionis/govern@v1
```

Gate a CI workflow so every outbound action is governed by the Decionis authority the workload trusts.

```yaml
- uses: decionis/govern@v1
```

— Exact action name and workflow path given, mechanism thin. product docs

THINNER COVERAGE BELOW

04 **Destination governance with operator-held TLS** NEW

30

Destination governance is now enforced from the interceptor, with TLS terminated under an authority the operator holds.

— No command or config surface named, brief description only. v0.3.2

05 **Scheduled shadow reporting with Compromised Principal Test**

NEW

30

Shadow reporting now runs on its own cadence, separate from live enforcement, with the Compromised Principal Test observable in the report.

— Feature named but no command or config detail given. v0.3.2

v0.3.2 NOTES ↗ LATEST STABLE

agent-safe-pipeline v0.3.2 adds destination governance with operator-held TLS, shadow report scheduling, Compromised Principal Test, and `agentsafe test --hosted`.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.3.2 https://github.com/decionis/agent-safe-pipeline.git
# already have the repo? check out this version:
$ git checkout v0.3.2
```

└──▷ TRY IT

Bootstrap a workspace's first governed action — run the hosted boundary test without any local policy configuration.

```
$ agentsafe test --hosted
```

Provision credentials during login in a fresh environment where no prior session exists.

```
$ agentsafe login --provision
```

Adds `agentsafe test --hosted` to run the boundary test with Decionis deciding, putting a workspace's first governed action one command away.

Adds `agentsafe login --provision` to provision credentials as part of the login flow.

> Adds destination governance from the interceptor, with TLS terminated under an authority the operator holds.
>
> Adds shadow reporting on its own cadence with the Compromised Principal Test observable in the report.

WAS THIS USEFUL? 👍 👎

Docker Sandboxes

Feedback ⓘ v0.45.0 NOTES ↗ RANK

Docker Sandboxes runs AI agents in isolated Docker environments with controlled filesystem, network, and tool access.

> Docker Sandboxes' v0.45.0 release introduces v3 OCI kits that split workloads and mixins into composable, dependency-aware packages, alongside new MCP gateway credential handling (custom headers, secret substitution, OAuth namespacing), a raft of CLI flags and JSON output fields, and several network-policy security fixes.

└──▷ **WHAT SHIPPED · 11 FEATURES** most completely described first ? what's the number?

01 Custom HTTP headers and secret substitution for MCP servers

NEW 95

Adds `--header 'Name: value'` to `sbx mcp add` for sending custom HTTP headers (e.g. `Authorization`, `Accept`) to remote MCP endpoints, repeatable per header. Credential headers are stored as named secrets via `sbx secret set mcp:<server>:<placeholder>`, keeping plaintext out of registrations and substituting values at connection time; `mcp:-prefixed` credential names are reserved in the host credential store for MCP gateway OAuth client secrets and custom header secrets, keeping them on the host and out of sandbox injection. `sbx mcp inspect <server>` checks header templates and whether their secrets are set without revealing resolved values, and `sbx secret rm mcp:<server>:<placeholder>` removes header secrets individually after a registration is removed.

Register a remote MCP server that authenticates via API key, storing the key in the host credential store so the gateway substitutes it at runtime.

```
$ sbx secret set mcp:acme:api-key
sbx mcp add acme --url https://mcp.acme.com/mcp \
  --header 'Authorization: Bearer ${api-key}' \
  --header 'Accept: application/json, text/event-stream'
sbx run claude --name acme-demo --static-mcp acme
```

Verify that all header templates for a registered MCP server have their secrets populated before launching a sandbox.

```
$ sbx mcp inspect acme
```

— Full command set given, mechanism for substitution explained clearly. product docs

02 **Custom registry auth endpoint for sandbox credentials** NEW

85

Adds `--registry-auth-endpoint` to trust a separate authentication host (e.g. `https://gitlab.example.com/jwt/auth`) when a self-hosted registry advertises an external realm in its `WWW-Authenticate: Bearer` challenge, used with `docker sandbox credentials add`.

Authenticate sandboxed agents against a self-hosted GitLab registry that delegates token issuance to a separate auth host.

```
$ docker sandbox credentials add \
  --sandbox my-app \
  --registry registry.example.com \
  --username "$GITLAB_USER" \
  --password "$GITLAB_PAT" \
  --registry-auth-endpoint https://gitlab.example.com/jwt/auth
```

— Exact flag, trigger condition, and runnable example given. product docs

03 **Custom model provider configuration** NEW

85

Adds a `model.providers` setting so `sbx run --model` can target any OpenAI- or Anthropic-compatible endpoint. Also exposes a model's supported thinking levels as selectable OpenCode variants (via Ctrl+T) through `sbx run opencode --model`.

Route `sbx run` to a self-hosted Anthropic-compatible endpoint instead of the default provider.

```
$ sbx run --model my-provider/claude-3-5-sonnet 'summarize this
repo'
```

— Config key and runnable command with target endpoint shown. v0.45.0

04 **V3 OCI kits with workload/mixin composition and kit sets** NEW

75

Introduces v3 kits: OCI-based packages with separate workload and mixin roles, dependency/compatibility declarations, and composition-order control. Kit sets let authors pin a workload and mixins, add lifecycle hooks, capabilities, instructions, and arguments, and publish the result as a single OCI reference. V3 kits can scope network access by HTTP method and path, declare install-phase network access and credential use, and specify where an agent reads shared skills; multiple OAuth-backed agents can now be composed in the same sandbox with credentials scoped to the kits that request them. V3 workloads and mixins must be used together and cannot be combined with v1 or v2 kits.

— Rich mechanism described but no exact command walkthrough. v0.45.0

05 **CLI flags and JSON output additions across commands** IMPROVED

75

Adds `--quiet` / `-q` to MCP server, secret, skill, template, volume, and policy-profile list commands for name-only output; `--created-via` filter to `sbx policy ls`, which now also shows how each rule was created; `--disable-http2` to `sbx mcp add` to disable HTTP/2 for a remote MCP transport; `--protocol` to local network allow rules so UDP can be selected explicitly alongside TCP (experimental outbound UDP now follows sandbox network policy); `--force` (and `--yes` / `-y` for `sbx kit builder history rm`) to skip confirmation prompts on destructive commands, with declining now returning a non-zero exit code; `created_at`, CPU limits, and memory limits to `sbx ls --json` (CPU/memory limits also appear in `sbx inspect --json`); and `sbx settings` to `sbx --help` and the CLI reference.

Allow outbound UDP to a specific host — for example, a custom DNS or VPN endpoint — while keeping TCP as the default for other rules.

```
$ sbx policy allow network udp://10.0.1.53:53 --protocol udp
```

Work around an MCP server whose HTTP/2 handling stalls long-lived streams by disabling HTTP/2 on registration.

```
$ sbx mcp add https://mcp.example.com/sse --disable-http2
```

— Many exact flags named, two with runnable examples. v0.45.0

06 **MCP catalog removal and stricter error handling** BREAKING

70

`sbx mcp catalog` has been removed; register a remote MCP server with `sbx mcp add` and then authorize it with `sbx mcp auth`. `sbx secret rm` now returns an error on stderr when the requested secret does not exist, and `sbx mcp rm` now returns an error when the requested MCP server is not registered.

— Names removed command and its exact replacement path. v0.45.0

07 **Sandbox listing filters and stop/resume behavior** IMPROVED

60

Adds an `until` filter to sandbox listing, accepting a duration, RFC 3339 timestamp, or Unix timestamp, to scope results by when a sandbox stopped. Documents automatic

stop/restart behavior: resuming a previously stopped sandbox with `sbx run --name <sandbox-name>` preserves files and config without recreating it.

Resume a previously stopped sandbox — files and config are intact — without recreating it from scratch.

```
$ sbx run --name <sandbox-name>
```

— Filter formats specified but lifecycle mechanics stay high-level. product docs

08 **Kit validation for API-key declarations** IMPROVED 60

`sbx kit validate` now rejects malformed API-key declarations, including invalid names, missing injection domains, invalid format placeholders, and Basic-auth usernames containing a colon, and warns on no-effect or out-of-allowlist declarations.

— Named checks but no example command run. v0.45.0

THINNER COVERAGE BELOW

09 **Network policy hostname and DNS lookup hardening** IMPROVED 55

Network policy now treats hostnames with a trailing dot as equivalent to their canonical form for routing, interception, and credential injection. Reverse-DNS lookups are now restricted to destination IPs already authorized by policy, closing a previous policy bypass.

— Describes the fix and its scope but no verification command. v0.45.0

10 **Secret management improvements** IMPROVED 50

Running `sbx secret rm` without a service now opens an interactive picker showing existing local secrets with their scope, type, and name. Adding, updating, or removing global service secrets now propagates to existing local sandboxes immediately without a restart.

— Behavior described, no command syntax to try it. v0.45.0

11 **Signed macOS app bundle distribution** IMPROVED 40

macOS distributions now ship as a single signed `Sbx.app` bundle, with Homebrew and tarball PATH installs working via a symlink into the bundle.

— States the packaging change without install steps. v0.45.0

└──▷ BREAKING ON UPGRADE

! `sbx mcp catalog` has been removed; register a remote MCP server with `sbx mcp add` and then authorize it with `sbx mcp auth`.

! `sbx secret rm` now returns an error on stderr when the requested secret does not exist.

! `sbx mcp rm` now returns an error when the requested MCP server is not registered.

! V3 workloads and mixins must be used together; they cannot be combined with v1 or v2 kits.

v0.45.0 NOTES ↗ LATEST STABLE

Docker Sandboxes v0.45.0 ships v3 OCI kits with workload/mixin composition, new model provider config, and expanded CLI controls.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch v0.45.0 https://github.com/docker/sbx-releases.git
# already have the repo? check out this version:
$ git checkout v0.45.0
```

└──▷ **TRY IT**

Route `sbx run` to a self-hosted Anthropic-compatible endpoint instead of the default provider.

```
$ sbx run --model my-provider/claude-3-5-sonnet 'summarize this repo'
```

Allow outbound UDP to a specific host — for example, a custom DNS or VPN endpoint — while keeping TCP as the default for other rules.

```
$ sbx policy allow network udp://10.0.1.53:53 --protocol udp
```

Work around an MCP server whose HTTP/2 handling stalls long-lived streams by disabling HTTP/2 on registration.

```
$ sbx mcp add https://mcp.example.com/sse --disable-http2
```

Adds `model.providers` setting so `sbx run --model` can target any OpenAI- or Anthropic-compatible endpoint.

Adds `sbx kit validate` rejection of malformed API-key declarations, including invalid names, missing injection domains, invalid format placeholders, and Basic-auth usernames containing a colon, with warnings for no-effect or out-of-allowlist declarations.

Adds `--quiet` / `-q` flag to MCP server, secret, skill, template, volume, and policy-profile list commands for name-only output.

Adds `--protocol` flag to local network allow rules so UDP can be selected explicitly alongside TCP (experimental outbound UDP now follows sandbox network policy).

Adds `--created-via` filter flag to `sbx policy ls`, which now also shows how each rule was created.

+14 more

└─▷ **BREAKING ON UPGRADE**

- ! `sbx mcp catalog` has been removed; register a remote MCP server with `sbx mcp add` and then authorize it with `sbx mcp auth`.
- ! `sbx secret rm` now returns an error on stderr when the requested secret does not exist.
- ! `sbx mcp rm` now returns an error when the requested MCP server is not registered.
- ! V3 workloads and mixins must be used together; they cannot be combined with v1 or v2 kits.

WAS THIS USEFUL? 👍 👎

vArmor

Feedback ⓘ v0.10.5 NOTES ↗ RANK

The vArmor Kubernetes runtime security platform provides AppArmor, Seccomp, and eBPF policy enforcement for workloads.

> vArmor's v0.10.5 release adds file-based SDS for hot-reloading MITM certificates and trust bundles, alongside several breaking changes covering audit sink upgrade ordering, HTTP method matching, proxy configuration regeneration, and MITM domain/Secret handling.

└──▷ **WHAT SHIPPED · 5 FEATURES** most completely described first ? what's the number?

01 **Literal HTTP method matching in policies** BREAKING 65

HTTP methods now match literally: policies using `get` to mean `GET` must explicitly use `GET`; rules relying on encoded path separators must be reviewed.

— Names exact values needing change in existing policies. v0.10.5

02 **MITM domain and Secret header value cleanup** BREAKING 65

Duplicate or whitespace-padded MITM domains and empty Secret-backed header values must be removed; Secret changes alone are not watched — a policy-triggered reconciliation is required to pick up updates.

— Explains the watch limitation and required workaround precisely. v0.10.5

03 **Audit sink upgrade ordering requirement** BREAKING 60

Agents and embedded Kata/VCI audit sinks must be upgraded before the Manager — older sinks reject the new audit event class introduced in this release.

— States exact upgrade order and reason for the break. v0.10.5

THINNER COVERAGE BELOW

04 **File-based SDS for MITM certificate hot-reload** NEW 55

Adds file-based SDS (Secret Discovery Service) for hot-reloading MITM certificates and upstream trust bundles without restarting the proxy.

— Describes mechanism but no config key or command given. v0.10.5

05 **Proxy config regeneration for SDS support** BREAKING 55

Proxy configuration must be regenerated and affected Pods recreated to apply SDS mounts, bootstrap settings, and container security-context fixes.

— Gives concrete migration steps but no exact commands. v0.10.5

└──▷ BREAKING ON UPGRADE

! Agents and embedded Kata/VCI audit sinks must be upgraded before the Manager — older sinks reject the new audit event class.

! HTTP methods now match literally: policies using `get` to mean `GET` must explicitly use `GET`; rules relying on encoded path separators must be reviewed.

! Proxy configuration must be regenerated and affected Pods recreated to apply SDS mounts, bootstrap settings, and container security-context fixes.

! Duplicate or whitespace-padded MITM domains and empty Secret-backed header values must be removed; Secret changes alone are not watched — a policy-triggered reconciliation is required.

v0.10.5 NOTES ↗ LATEST STABLE

vArmor v0.10.5 adds file-based SDS for hot-reloading MITM certificates and upstream trust bundles.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.10.5 https://github.com/bytedance/vArmor.git
# already have the repo? check out this version:
$ git checkout v0.10.5
```

└──▷ BREAKING ON UPGRADE

! Agents and embedded Kata/VCI audit sinks must be upgraded before the Manager — older sinks reject the new audit event class.

! HTTP methods now match literally: policies using `get` to mean `GET` must explicitly use `GET`; rules relying on encoded path separators must be reviewed.

! Proxy configuration must be regenerated and affected Pods recreated to apply SDS mounts, bootstrap settings, and container security-context fixes.

! Duplicate or whitespace-padded MITM domains and empty Secret-backed header values must be removed; Secret changes alone are not watched — a policy-triggered reconciliation is required.

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Open-Weight Models

Alibaba Qwen Models

Feedback ⓘ qwen3.8-omni-flash-realtime

NOTES ↗ RANK

Qwen is Alibaba's language and multimodal model family, including coding, vision, image, speech, and reasoning models.

> Alibaba introduced Qwen3.8-Omni-Flash-Realtime, a new real-time multimodal model supporting multichannel audio, video aggregation, remote tool integration, and multiple connection protocols.

└──▷ **WHAT SHIPPED · 1 FEATURE** most completely described first ? what's the number?

01 Qwen3.8-Omni-Flash-Realtime real-time multimodal model NEW

65

New model supports real-time audio and video interaction with text and audio output. It adds multichannel audio support and video aggregation for real-time interactions, integrates remote MCP (Model Context Protocol) tools, and can be accessed via `WebSocket`, `WebRTC`, and `AOQ` protocols.

— Names protocols and MCP integration but no usage example or endpoint details.

qwen3.8-omni-flash-realtime

qwen3.8-omni-flash-realtime NOTES ↗ LATEST STABLE

Qwen3.8-Omni-Flash-Realtime adds multichannel audio, video aggregation, remote MCP tools, and WebRTC/AOQ access.

Supports real-time audio and video interaction with text and audio output.

Adds multichannel audio support for real-time interactions.

Adds video aggregation for real-time interactions.

Adds remote MCP (Model Context Protocol) tools integration.

Supports WebSocket, WebRTC, and AOQ access protocols.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · September 21, 2026

Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue