THE AI TOOLCHAIN

NO. 076

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED OCTOBER 28, 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 477 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 116 tools

▾ // THE BRIEF HIDE

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. OpenSandbox pauses idle sandboxes and wakes them on first access with a defined retry contract, so long-agent environments stop billing between turns. Jan closes its local

vim ? help

API server to untrusted origins by default. Axolotl writes sharded LoRA checkpoints without gathering to rank 0, taking a stall out of every save in a long run.

DEPLOY

Keep an agent sandbox alive for days without paying for the idle hours

OpenSandbox auto-pauses fast sandboxes and wakes them when a request arrives, with a documented wake-retry contract so callers know how to handle the cold moment; OpenShell adds suspend/resume for the same shape of workload. Previously the choice was keeping a box warm indefinitely or rebuilding state on every session.

opensandbox · OpenShell

BUILD

Save sharded LoRA checkpoints without the rank-0 gather that stalls every node

Axolotl's new sharded LoRA/QLoRA checkpoint format skips the all-gather to rank 0 and mirrors checkpoints in the background, so saves stop blocking the training step on large multi-node runs. compressed-tensors adds MXFP6 and MXFP4xFP8 plus a fused NVFP4 kernel and faster 4-bit unpacking, cutting the time between a trained model and a quantized one that serves.

Axolotl · compressed-tensors

GOVERN

Stop a local or hosted agent server from answering calls it should never see

Jan adds Trusted Hosts and origin checking to its local API server plus CORS enforcement and restricted environment variables for Claude Code — a local inference port is reachable from any page the user visits unless something checks who is asking. Arcade withholds secrets per scope for API-registered workers and redacts them in tool responses and logs; gridctl adds external secret providers and per-tool execution policy blocks.

Jan · Arcade

OFFENSIVE SECURITY

◆ Exploitation & C2

xalgorix

Feedback ⓘ v4.6.173 NOTES ↗ RANK

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

> xalgorix v4.6.173 adds parser support for DSML-namespaced tool-call markup, letting agents dispatch these calls natively instead of treating them as plain text.

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

01 **DSML-namespaced tool-call parsing** NEW 60

The agent now supports DSML-namespaced tool-call markup, including `<｜DSML｜invoke name="…">` and `<｜DSML｜parameter …>` elements, dispatching them like any other tool call instead of treating them as plain text.

— Names exact markup syntax but no usage example or config key v4.6.173

WAS THIS USEFUL? 👍 👎

BUILD

◆ AI Coding Agents

zot

Feedback ⓘ v0.4.53 NOTES ↗ RANK

zot manages coding agent workflows from the command line.

> zot v0.4.53 adds built-in MCP (Model Context Protocol) support and lets the TUI attach to an existing continuous conversation.

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

01 **Built-in MCP support** NEW 35

zot adds built-in MCP (Model Context Protocol) support via feat(mcp).

— Names the protocol and commit tag but no usage detail v0.4.53

02 **TUI attaches to existing conversation** NEW 35

zot adds feat(continuous) capability allowing the TUI to attach to an existing conversation.

— Names the commit tag but lacks mechanism or steps v0.4.53

WAS THIS USEFUL? 👍 👎

yetone magpie

Feedback ⓘ v0.1.1850 CODE ↗ RANK

Magpie is a macOS menu-bar app, TUI, and CLI that lists every AI coding agent on the machine and switches each one's model, such as Codex on DeepSeek or Claude Code on Kimi, by rewriting only that key in the agent's own config.

> Magpie shipped context-aware Chinese phone number redaction, fixed Claude Desktop connection probes to hit the gateway's default model, made the plugin dist index resilient to stale jsDelivr caches via a GitHub fallback and China mirror, and forced Bun to use the OS certificate store to fix TLS failures behind corporate proxies.

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

01 **System CA trust store for Bun runtime** NEW 78

Sets `NODE_USE_SYSTEM_CA=1` in the Bun environment so Bun reads the OS's own certificate store (Windows ROOT/CA/TrustedPeople stores, macOS keychains) instead of its bundled roots, resolving `UNABLE_TO_GET_ISSUER_CERT_LOCALLY` and 'self signed certificate in certificate chain' errors behind corporate proxies, VPNs, and antivirus HTTPS inspection.

— Names exact env var and specific errors it resolves v0.1.1826

02 **Context-aware Chinese phone number redaction** NEW 73

The `redact` subsystem now masks eleven-digit strings starting 13–19 only when context confirms they are phone numbers — via a `+86` / `86` prefix, a field name containing `phone`, `mobile`, `tel`, `cell`, `contact`, `whatsapp`, `手机`, or `电话`, or prose lacking an ID-indicating word (`order`, `id`, `no.`, `#`, `订单号`, `编号`, `单号`, `流水号`, `工单`) immediately before the digits.

— Clear mechanism and named rules, but it's automatic with no user action v0.1.1850

03 **Plugin dist index resilience via GitHub and China mirror**

IMPROVED 73

Adds `readDistIndex` to fetch the community dist branch `index.json` directly from GitHub alongside the jsDelivr mirror, so dist-only plugins (never published to npm) get their true latest version even when jsDelivr's edge cache is stale. Also exposes `source.ChinaFirst` to return the mirror URL for a resource when the `国内镜像` switch is on, so the dist index fetch consults the China mirror before falling back to GitHub.

— Names function, file, and a user-facing mirror switch v0.1.1829

04 **Claude Desktop probes target gateway default model** IMPROVED

62

Claude Desktop startup and `Test connection` probes (`desktopProbe`) now route to the gateway's default model (`desktopDefault`, the first listed model) instead of whichever model the last session ran on, preventing probes from landing on a stale or wrong-provider model.

— Named internals explained but it's an automatic fix, nothing to invoke `v0.1.1847`

WAS THIS USEFUL? 👍 👎

ogcode

Feedback ⓘ `v0.53.1` NOTES ↗ RANK

An open-source, local-first AI teammate and agent operating environment for turning intent into completed work—with tools, memory, permissions, and human oversight.

> ogcode's single release in this window refines the Models settings UI, scoping each vendor slot to its own endpoint's models and organizing providers into separate tabs.

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

01 Per-vendor model filtering and tabs in Models settings

IMPROVED

50

Vendor slots in Models settings now list only the models belonging to their own endpoint, reducing noise when multiple providers are configured. Models settings also gains one tab per vendor for clearer per-provider model management.

— Clear UI change and scope, but no exact navigation path given `v0.53.1`

WAS THIS USEFUL? 👍 👎

kane-cli

Feedback ⓘ 0.10.0 NOTES ↗ RANK

Kane CLI by TestMu AI (Formerly LambdaTest)

> kane-cli 0.10.0 moves exploration execution to HyperExecute by default with new run controls and policy gates, adds a Test Manager policy CLI wired into evidence verdicts, makes agent-friendly NDJSON output universal, and ships a live React Flow map for watching explorations in real time.

└──▷ **WHAT SHIPPED · 9 FEATURES** most completely described first ? what's the number?

01 Policy gates for explore run NEW

95

`explore run --policy read_only|mutating` controls what a run can do; mutating is now the default, while read-only blocks writes at the route and browser level and refuses secret controls. `--allow-post` names read-only POST endpoints the gate should pass through, or `*` to pass every POST to the run's own domains.

— Explains default behaviour, blocking mechanism and the escape-hatch flag. 0.10.0

02 Unattended assurance-job for design chain NEW

95

`kane-cli explore assurance-job` runs the full design chain unattended, returning outcome codes `MARK_REFUSED`, `PROJECT_NOT_MARKED`, `LATTICE_UNAVAILABLE`, `INGEST_FAILED`, and `INPUT_INVALID`. The `KANE_AJ_CURSOR_FROM` environment variable bounds the job's ingestion cut from below.

Run the full design chain unattended in CI, bounding the job's ingestion cut to a known cursor.

```
$ KANE_AJ_CURSOR_FROM=2026-10-01T00:00:00Z kane-cli explore assurance-job
```

— Lists all outcome codes, the env var, and a runnable CI example. 0.10.0

03 Remote exploration execution on HyperExecute NEW

90

`explore run` now executes remotely on HyperExecute by default, with `--claim-run` used by the VM to claim a queued run, `--focus <screen>` to start exploration from one screen, `--depth-caps` for per-depth lease caps, and `--max-parallel N` to bound concurrency. `explore stop <id>` aborts a queued run or cleanly stops one already running.

— Names every flag and the stop command with exact syntax. 0.10.0

04 Test Manager policy CLI and evidence verdicts NEW

90

`kane-cli policy` adds subcommands to list, create and inspect Test Manager policies, plus `kane-cli policy metrics-list` for the metric catalog. A new `--policy-id` flag on `evidence merge`, `testrun run`, and `testmd run` judges a sealed evidence pack against a policy, and `evidence verdict` and `evidence lint-policy` read and lint policies over the evidence-cli verdict API.

Judge a sealed merged evidence pack against a Test Manager policy to gate a release.

```
$ kane-cli evidence merge --policy-id pol-abc123
```

— Enumerates every subcommand and flag with a runnable example. 0.10.0

05 **Agent NDJSON output across commands** NEW 73

A `--agent` flag is now available universally across every command for NDJSON output, with `--mode agent` as an alias on `run` and `testmd run`, and `--json` added to `extract`, `design tests`, and `maintain reconcile`.

— Names every flag and the commands it applies to. 0.10.0

06 **Live exploration map in explore view** NEW 73

`explore view` renders a live React Flow map of an exploration with zoom, search, a legend, a findings drawer with filters, and map badges. It streams the live browser through the exploration controller at up to 15 fps per browser, accessible via the run card's Watch button.

— Rich feature detail but entry point is UI navigation, not a command. 0.10.0

07 **Remote browser driving via LambdaTest CDP** NEW 70

`explore --browser remote` drives LambdaTest Chrome over CDP, running up to 20 cloud browsers simultaneously.

— Names the flag, protocol and concurrency limit. 0.10.0

08 **Agent-stream review decisions in context review** NEW 70

`context review` gains `--approve`, `--skip`, and `--defer` flags plus `--mode override` for agent-stream decisions in a linked folder.

— Exact flags named but mechanism beyond that is unexplained. 0.10.0

THINNER COVERAGE BELOW

09 **Session search across run prefix pages** NEW 55

`explore sessions --run <prefix>` matches a run prefix across every page of session results.

— Thin single-flag addition with no further mechanism given. 0.10.0

WAS THIS USEFUL? 👍 👎

claude-mem

Feedback ⓘ v15.4.0 NOTES ↗ RANK

claude-mem gives coding agents persistent memory by capturing session activity, compressing it with a model, and injecting relevant context into later sessions.

> claude-mem v15.4.0 introduces a My Work focus-area organizer that automatically files and suggests groupings for requests, overhauls the automatic refresh pipeline with AI-driven join scoring, and opens a live decisions API for Pro subscribers.

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

01 Automatic request filing and join scoring IMPROVED

80

Overhauls the automatic refresh pipeline: files each new conversation to an open request via a `request_id` code link before any grouping agent runs; adds Jev-powered join scoring on CMEM Pro or with an `OpenRouter` key, auto-joining conversations scoring 0.8+ and surfacing a `Join / Not this` suggestion for scores between 0.5 and 0.8; adds a Jev focus-area suggestion per request before the Focus agent runs, seeding automatic placement once enough user moves exist to calibrate a cutoff; and replaces the Align step with a Focus step that places every request and proposes new areas from Other or multi-ask requests.

– Thresholds and mechanism named, but no direct command to invoke it. v15.4.0

02 My Work focus-area organizer NEW

78

Launches the `MY FOCUS` workflow: a rail lists focus areas in priority order with a needs-you count, filing every request into exactly one area or Other (rendered as a chip row on phones); per-area pages show a goal sentence, combined progress, user quotes, unstarted asks, and a link to Past focus. A Suggested section surfaces new focus areas automatically with Keep, Rename, Merge, and Dismiss actions that sync across computers, and a Move to… control on request detail cards lets users manually reassign a request to a different focus area.

– Names every UI control and action but gives no API/CLI entry point. v15.4.0

03 Live decisions API gateway for Pro NEW

65

Adds `/api/inference/v1/decisions` gateway live for CMEM Pro subscribers.

– Exact endpoint path named but no method or payload details given. v15.4.0

THINNER COVERAGE BELOW

04 Cross-project-root request persistence NEW

50

Enables a request to retain its conversations across project roots when its decision names a project root, with later refreshes still able to add to it.

— Mechanism described in prose without a named config or command. v15.4.0

---

05 **My Goals page retirement** BREAKING 30

The My Goals page is retired; its content now lives inside the per-area pages under My Work.

— States the change but no migration steps beyond page relocation. v15.4.0

└──▷ **BREAKING ON UPGRADE**

! The My Goals page is retired; its content now lives inside the per-area pages under My Work.

WAS THIS USEFUL? 👍 👎

Token Optimizer

Feedback ⓘ v5.13.60 NOTES ↗ RANK

Token Optimizer is a coding-agent plugin that trims wasted context, checkpoints work across sessions and compactions, and reports token spend on a live dashboard.

> Token Optimizer brought its full feature set to Hermes as a plugin (v0.2.0) with compressed output, checkpoints, and secret masking, while tightening Hermes profile isolation and fixing hook reliability across Codex, Copilot CLI, and Claude Code.

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

01 **Hermes plugin support with full feature set** NEW 90

Adds the full Token Optimizer feature set to Hermes via `install.sh --hermes`, delivering plugin 0.2.0 with compressed terminal output, smart re-reads, checkpoints that survive compaction, pick-up-where-you-left-off, and secret masking.

Install the Hermes plugin to get compressed output, checkpoints, and secret masking on that platform for the first time.

```
$ install.sh --hermes
```

— Names exact install flag, plugin version, and five capabilities v5.13.59

02 **Hermes profile isolation and reinstall auditing** IMPROVED 80

`hermes-doctor` now audits every Hermes profile to identify which ones need a reinstall, and each profile's sessions, checkpoints, and saved output are isolated into its own dashboard when running a multiplexed gateway or cron on a second profile — re-run `hermes-install` in each profile, then restart Hermes.

After adding a second Hermes profile (e.g. for a cron job), check which profiles still need a reinstall before restarting.

```
$ hermes-doctor
```

— Names both commands and the exact remediation steps v5.13.60

03 **Hook reliability fixes for Codex, Copilot CLI, Claude Code** IMPROVED 70

Codex hooks now run correctly in both PowerShell and cmd.exe, including user-folder paths containing spaces, `$`, `%`, or `!`. Copilot CLI hooks now fire correctly (restore bash output compression with `copilot-install`), and Claude Code hooks remain active after a plugin update.

Restore Copilot CLI bash output compression after upgrading to v5.13.59.

```
$ copilot-install
```

— Names affected shells, special characters, and a restore command v5.13.59

THINNER COVERAGE BELOW

04 **Checkpoints capture real task description** IMPROVED 45

Checkpoints now record the real task description in any language, rather than system notes or a `/compact` line.

— Clear before/after but no user action involved v5.13.59

05 **Usage trend labels flag cache reads** IMPROVED 25

Usage trend labels now indicate which numbers include cache reads.

— Bare description with no mechanism or scope given v5.13.60

WAS THIS USEFUL? 👍 👎

Letta Code

Feedback ⓘ v0.34.26 NOTES ↗ RANK

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

> Letta Code added a new Cursor-based coding worker agent type, webhook routing to connected computers, and a hidden memory subagent for automatic repair.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 **Cursor coding workers agent type** NEW 35

Adds Cursor coding workers as a new agent type, shipped via `feat(agent): add Cursor coding workers`.

— Names the commit but not usage or configuration v0.34.26

02 **Webhook skill targets connected computer** NEW 35

Lets a webhook skill target a connected computer, enabling remote-computer routing from webhook-triggered workflows.

— Explains behaviour but no concrete config or endpoint v0.34.26

03 **Hidden memory subagent for automatic repair** NEW 20

Introduces a hidden memory subagent that performs automatic repair, as mentioned in the release summary.

— Only named in summary with no mechanism detail v0.34.26

WAS THIS USEFUL? 👍 👎

OpenQodex

Feedback ⓘ openqodex@0.16.0 NOTES ↗ RANK

Open source code review that runs in your coding agent before you push. Scanners plus a separate AI reviewer, on your own login, no key.

> OpenQodex 0.16.0 deepens its code graph with JUnit test-to-code linking, more precise Java call resolution, multi-language receiver tracking, and separates possible issues from the review verdict.

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

01 Java generic call dispatch resolution IMPROVED 65

Java call dispatch through generic type parameters is improved, so chains like Box<Repo>.get().find() now correctly bind to `Repo.find` in the code graph.

– Concrete before/after example but no runnable command. openqodex@0.16.0

THINNER COVERAGE BELOW

02 Pattern variable tracking after negated guards IMPROVED 55

The Java code graph now follows pattern variables after negated guards and negated while loops, improving accuracy of flow tracking in these constructs.

– Names the mechanism but gives no usage example. openqodex@0.16.0

03 Typed Go and TypeScript receiver chaining NEW 55

The code graph is extended to follow declared method results on typed Go and TypeScript receivers, retaining chained callers across these languages.

– Names languages and scope but no concrete example or command. openqodex@0.16.0

04 JUnit test-to-code linking in code graph NEW 50

The code graph now links JUnit 4 and JUnit 5 test methods to the Java code they call, connecting tests directly to the implementation they exercise.

– Describes what changed but no mechanism or usage path given. openqodex@0.16.0

05 Possible issues tracked separately from verdict IMPROVED 45

Possible issues are now shown separately in results and excluded from the verdict count, so they no longer affect the overall pass/fail verdict.

– States behaviour change but no detail on how it surfaces to users. openqodex@0.16.0

WAS THIS USEFUL?  

OpenAI Codex CLI

Feedback ⓘ rust-v0.166.0-alpha.4 NOTES ↗

CODE ↗ RANK

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

> This window's three alpha releases of Codex CLI add opt-in protocol discovery for local MCP servers, a standalone `codex-exec-server` binary, granular approval catalog fields with byte-limited templates, and bearer-token defaults for gateway OAuth delivery, alongside smaller TUI, catalog and safety-messaging updates.

└──▷ **WHAT SHIPPED · 12 FEATURES** most completely described first ? what's the number?

01 **Granular approval catalog fields** NEW 80

Adds optional `approvals.granular` catalog fields — including policy introduction, prompted category list, and rejected category list — for granular approval templates, with a 500-byte limit and required `{{ policy }}` and `{{ categories }}` markers.

— Names config fields, limit, and required markers precisely rust-v0.166.0-alpha.2

02 **Bearer-token default for gateway OAuth delivery**

NEW SUPPORTED PLATFORM IMPROVED 80

Allows omitting `gateway_oauth.delivery` to default gateway OAuth token delivery to `Authorization: Bearer …`, with validation that `model_catalog_url` shares the scheme, host, and port of `base_url`.

— Names exact config keys and validation rule rust-v0.166.0-alpha.2

03 **Concise plugin selection hints flag** NEW 73

Adds `features.simplified_selection_hints` flag (disabled by default) to replace verbose plugin capability instructions with a concise developer message identifying the selected plugin.

— Named flag with default state is directly actionable rust-v0.166.0-alpha.2

THINNER COVERAGE BELOW

04 **Standalone codex-exec-server binary** NEW 55

Adds a standalone `codex-exec-server` executable, now shipped in both the CLI and app-server packages.

— Names concrete binary and packages but no usage detail rust-v0.166.0-alpha.4

05 **ChatGPT workspace shown in /status** IMPROVED 55

Shows the selected ChatGPT workspace in the `/status` command in the TUI.

— Exact command named but limited detail beyond display `rust-v0.166.0-alpha.2`

06 **Opt-in protocol discovery for MCP servers** NEW 45

Codex CLI now supports opt-in protocol discovery for local stdio MCP servers, allowing negotiation of supported protocol features on a per-server basis.

— Names mechanism but no flag or command given `rust-v0.166.0-alpha.5`

07 **Tool and approval catalog customization** NEW 45

Apps tool catalogs can now operate without shared snapshots, enabling dynamic tool-catalog replacement between model steps, and model catalogs can customize approval instruction headings.

— Describes behavior but no config keys or commands named `rust-v0.166.0-alpha.4`

08 **Safety policy block explanations with recourse** IMPROVED 45

Explains safety policy blocks with turn-scoped recourse in the TUI and offers safer rewrites after safety blocks.

— Describes behavior but no command or config surface `rust-v0.166.0-alpha.2`

09 **TUI math rendering improvements** IMPROVED 40

Renders integral limits and parenthesized matrices in TUI math display.

— Specific rendering cases named, no further mechanism `rust-v0.166.0-alpha.2`

10 **Experimental GCP account setup** NEW INTEGRATION NEW 35

Adds experimental GCP account setup to the app server.

— Thin mention of experimental feature, no steps given `rust-v0.166.0-alpha.2`

11 **Filesystem-free classification of patch commands** NEW 30

Adds filesystem-free classification of intercepted patch commands.

— One-line description with no mechanism detail `rust-v0.166.0-alpha.4`

12 **Literature search in web queries** NEW 25

Codex CLI web queries now support literature search.

— Bare mention with no further detail `rust-v0.166.0-alpha.4`

`rust-v0.165.0` NOTES ↗ CODE ↗ LATEST STABLE

Adds experimental attached thread forks to the app-server API, enables reasoning effort overrides by default, and enables memory v2 for fresh profiles.

└──▷ **GET THIS VERSION** ▾

Homebrew

```
$ brew install --cask codex
# already installed? upgrade to this version:
$ brew upgrade --cask codex
```

git

```
$ git clone --branch rust-v0.165.0
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.165.0
```

Homebrew lists this version · checked 28 Oct **Wrong command?**

Adds experimental attached thread forks to the app-server API.

Enables reasoning effort changes during a conversation by default for supported models.

Enables memory v2 by default for fresh profiles and adds support for importing external-agent memories into it.

Adds a repository filter to the session picker and directory grouping to the agent command center for easier navigation across linked worktrees.

WAS THIS USEFUL? 👍 👎

GitHub Copilot CLI

Feedback ⓘ v1.0.105 NOTES ↗ RANK

GitHub Copilot CLI is a terminal-based coding agent that can plan, write, and test code in local repositories.

> GitHub Copilot CLI now lets users select the Copilot Fusion High or Max models directly from long-context sessions.

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

01 **Model selection for Copilot Fusion High and Max** NEW

55

Long-context sessions can now select the Copilot Fusion High or Max models via the `/model` command.

— Names the exact command and model options, but no further mechanism given. v1.0.105

WAS THIS USEFUL? 👍 👎

coffee-cli

Feedback ⓘ v3.7.7 NOTES ↗ RANK

coffee-cli is a clean AI terminal workspace with quick input, folder switching, copy & paste, and image pasting.

> Coffee CLI's latest releases focus on collaborative child-session management, adding lifecycle controls, automatic result handling, localized search, and polished UI feedback for multi-session workflows.

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

01 Child-session lifecycle controls for collaboration NEW

63

Child sessions created during collaboration workflows can now be resumed to their original conversation history, removed individually, promoted into independent standalone sessions, or retained after task completion. Closing a parent session now offers a choice to end the whole group or keep its children running, and sessions can also be closed by group.

— Describes several concrete lifecycle behaviors but no exact commands or flags v3.7.7 v3.7.6

THINNER COVERAGE BELOW

02 Child session tabs, animations and status indicators IMPROVED

45

Child tabs now use themed styling with subtle entrance animations that play only for newly created children, plus two-level hover previews. Completion events are silent while summary and approval prompts retain their sounds. Tab status indicators were unified into green, orange, and blue states, with a working-sidebar spinner added for in-progress sessions.

— Good behavioral detail but purely visual/UX, no actionable surface v3.7.7 v3.7.6

03 Configuration-panel state persistence IMPROVED

40

Configuration-panel state now persists across orchestration, computer-control, and browser-control views, with background refresh added to reduce layout shifts.

— Names the affected views but no exact UI path or config key v3.7.7

04 Localized collaboration tool search IMPROVED

37

Collaboration tool search results are now localized across languages while raw argument and result details are preserved in the original language when expanded.

— Explains behavior clearly but no config or command surface given v3.7.7 v3.7.6

05 **Automatic result collection and duplicate-dispatch prevention**

NEW

30

Collaboration sessions now automatically collect results from child sessions and prevent duplicate dispatches.

— Names the mechanism but gives no detail on how it operates v3.7.7

WAS THIS USEFUL? 👍 👎

opencode-telegram-bot

Feedback ⓘ v0.26.6 NOTES ↗

RANK

opencode-telegram-bot is a Telegram client for OpenCode that lets you run and monitor AI coding tasks from your phone while they run locally.

> opencode-telegram-bot v0.26.6 adds automatic reconnection when the OpenCode V2 server moves ports, plus a condensed one-line progress summary in compact mode.

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

01 **Auto-reconnect to relocated OpenCode V2 server** IMPROVED

65

Adds automatic reconnection in `/opencode_start` and auto-restart when OpenCode V2 re-registers its background server on a new port after a self-update, so the bot keeps working without manual intervention.

— Names the command and trigger condition clearly. v0.26.6

02 **Compact one-line progress summary** IMPROVED

62

In compact mode, finished progress messages now collapse into a one-line summary showing only non-zero counts, e.g. `✅ commands: 3, reads: 4, changed files: 2`, and the message is removed entirely when there is nothing to count.

— Concrete example format shown, but no config key to toggle it. v0.26.6

WAS THIS USEFUL? 👍 👎

pstack-claude

Feedback ⓘ v0.9.106 NOTES ↗ RANK

pstack-claude is a port of Poteto's pstack agent framework to Claude Code, Codex, Copilot, Pi, and other AI coding assistants.

> pstack-claude added a new Autopilot-stack mode alongside shared blocking rules, rewrote `/statechart` onto a TypeScript compiler, reworked `/interrogate` into adversarial repro confirmation, introduced citation-backed `HOLD` logic in `/no-comments`, and added a Pi override to avoid tool-name conflicts.

└──▷ **WHAT SHIPPED · 9 FEATURES** most completely described first ? what's the number?

01 /interrogate reworked to adversarial repro confirmation

75

BREAKING

Reworks `/interrogate` to send adversaries to break a change or design, keeping only findings that a rerun of their command confirms, and it now reads the `interrogate reviewers` line in place of `arena runners`, reusing arena's runner step. It no longer uses a reviewer table, reviewer prompt, rubric, lead-judgment, or code-quality references; its reviewer-based review model is replaced by adversarial repro confirmation. A subagent that opens a PR no longer runs `/interrogate`; it runs only `/deslop` and `/no-comments`.

— Full mechanism, config line, and removed/replaced behaviors named v0.9.105

02 Citation-backed HOLD state in /no-comments NEW

70

Adds a `HOLD` state for comments that state a rule of a dependency, platform, vendor, or protocol — the agent must cite the page or file that refutes the rule before a kill is accepted; if it cannot, the comment stays as `HOLD` naming the rule and where to check it. Five rule pins in `tests/skill-rules.test.mjs` enforce the new citation and `HOLD` requirements, failing if a sync drops either policy fork.

— Names exact state, enforcement mechanism and test file v0.9.106

03 Autopilot-stack mode and routing rules NEW

70

Adds Autopilot-stack, a new autopilot variant that stacks a PR only on a PR it builds on, without merging, complementing the existing Autopilot-full mode. Plans now automatically route to Autopilot-stack when a PR builds on another PR and to Autopilot-full when every PR stands alone, and both modes adopt a new blocking rule where a PR blocks only for a problem shown by a failing run or a wrong doc line, with a problem already present on the PR's base not blocking it.

— Names both modes, routing logic and blocking rule v0.9.105

04 **/statechart rewritten on TypeScript compiler** BREAKING

65

Rewrites `/statechart` to run on a TypeScript compiler instead of 21 `.mjs` scripts; it now checks a model and draws it. It no longer generates a TLA+ skeleton from an agreed chart and no longer audits code against a spec — those capabilities are removed.

— Describes new mechanism and names removed capabilities v0.9.105

05 **Pi ask-tool override for conflicting extensions** NEW

60

Adds `pi ask tool: off` to the Pi override sheet, letting the extension skip registering `ask_user_question` and leaving the name to another extension such as a dedicated ask-user-question package. Updates `/setup-pstack` on Pi to automatically add `pi ask tool: off` to the override sheet when the user runs a conflicting extension.

— Names exact config override and setup command v0.9.104

THINNER COVERAGE BELOW

06 **Per-report failure isolation in /no-comments** IMPROVED

45

A rejected second report in `/no-comments` now fails only for that report's findings, not for the entire run — accepted findings still proceed through steps 3 to 6.

— Clear before/after but limited mechanism detail v0.9.106

07 **Arena cross-judge and runner seat selection** IMPROVED

45

Arena's cross-judge now comes only from a model family other than the parent's, and arena drops or replaces a runner seat the Agent tool rejects.

— States rule but little further mechanism v0.9.105

08 **/poteto-mode goal restatement and background execution** IMPROVED

45

`/poteto-mode` now restates the user's goal in its first message, briefs, PRs, and final reports, and runs steers and resumes of running agents in the background with model fallback chaining.

— Names command and new behaviors, limited mechanism depth v0.9.105

09 **New substitution and denylist rules** NEW

35

Adds new substitution rules covering arena's default-pool sentences and a Task error, with the denylist rejecting bare Task error and 're-reads both playbooks from trunk'.

— Named phrases but unclear mechanism or surface v0.9.105

└──▷ BREAKING ON UPGRADE

! `/statechart` no longer generates a TLA+ skeleton from an agreed chart and no longer audits code against a spec; those capabilities are removed.

! `/interrogate` no longer uses a reviewer table, reviewer prompt, rubric, lead-judgment, or code-quality references; its reviewer-based review model is replaced by adversarial repro confirmation.

! A subagent that opens a PR no longer runs `/interrogate`; it runs only `/deslop` and `/no-comments`.

WAS THIS USEFUL?  

ai-rulez

Feedback ⓘ v5.8.0 NOTES ↗ RANK

Author, generate, validate and govern AI agent rules, skills, plugins and MCP configuration across 52 coding assistants. Review with verified model evidence, evaluate skills, lock and sign dependencies, and publish bundles.

> ai-rulez v5.8.0 adds cross-agent plugin configuration that writes into Claude Code, Codex and OpenCode settings, ships two new GitHub Actions for governance gating and dependency updates, exposes lock/update operations as MCP tools, and adds scoped-fallback grouping plus a frontmatter exclusion flag, while deprecating the `skillspector` lint profile shortcut.

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

01 Cross-agent plugin management via `[[enabled_plugins]]` NEW

95

Adds a `[[enabled_plugins]]` config block with `id`, `marketplace.name`, `marketplace.source`, `marketplace.type`, and `opencode` keys. It writes `enabledPlugins` / `extraKnownMarketplaces` to `.claude/settings.json`, a `[plugins."<id>"]` table with `enabled = true` to `.codex/config.toml`, and the npm package to the `plugin` array of `opencode.json`. `generate --user` writes the same entries to user-level files; accepted marketplace sources are `github`, `git`, `url`, `directory`.

Enable a third-party marketplace plugin for Claude Code, Codex and OpenCode from a single config declaration.

toml

```toml
# In .ai-rulez/config.toml
[[enabled_plugins]]
id = "basemind@basemind"
opencode = "@basemind/opencode-plugin"

[enabled_plugins.marketplace]
name = "basemind"
source = "github"
type = "marketplace"
```

— full config schema, generated file targets, and runnable example v5.8.0

02 `check@v5` GitHub Action for governance gating NEW

95

Adds the `Goldziher/ai-rulez/check@v5` GitHub Actions sub-action, which runs the full governance gate — `migrate okf --check`, `validate --strict --no-local`, `lock --check`, `generate`, and `generate --check` with `--strict-config --no-local` — with inputs `version`, `config-directories`, `expected-config-count`, `plugin`, `strict`, `local`, and `token-budget`.

Run the full governance gate on every config root in CI without per-release edits to the workflow.

```yaml
- uses: Goldziher/ai-rulez/check@v5
  with:
    version: latest
    strict: true
    expected-config-count: 3
    token-budget: 8000
```

— exact gate commands, inputs, and runnable workflow example `v5.8.0`

---

03 **Removal of `skillspector` lint profile shortcut** BREAKING

95

`profile = "skillspector"` in `[[lint.external]]` is no longer accepted; configuration fails to load with an error listing allowed values `agnix`, `claude-plugin-validate`, `cisco-skill-scanner`, `snyk-agent-scan`. It must be replaced with an explicit entry using `command = ["skillspector", "scan", "--no-llm", "--format", "sarif", "{skill_dirs}"]`, `format = "sarif"`, `inputs = ["skills"]`, `egress = false`, `batch_size = 1`, `version = ">=2,<3"`, and `message_properties = ["evidence", "snippet"]`; the profile keys `required_flags`, `offline_notices`, and the profile-level default for `message_properties` are removed.

— exact migration config and failing error message given `v5.8.0`

---

04 **`update@v5` GitHub Action for dependency updates** NEW

90

Adds the `Goldziher/ai-rulez/update@v5` GitHub Actions sub-action, which runs `ai-rulez lock --outdated --format json`, `ai-rulez update`, `generate --no-local`, and `lock --content-only`, then commits and opens or refreshes a PR from branch `ai-rulez/update`; inputs include `version`, `names`, `kind`, `direct`, `branch`, `token`, and `working-directory`.

— exact command sequence and PR branch named, no usage example `v5.8.0`

---

05 **`fallback: false` frontmatter to exclude items** NEW

76

Adds a `fallback: false` frontmatter key for rule and context files to exclude a glob-scoped item from the folderless-preset fallback entirely; `validate` rejects non-boolean values with error code `AR304`.

— exact key and validation error code named, no example v5.8.0

---

06 **New MCP tools for lock and update management** NEW 73

Adds MCP tools `lock_outdated`, `update_plan`, `update`, and `lock_diff`, mirroring `lock --outdated`, `update --dry-run`, `update`, and `lock --diff`; setting `[mcp] read_only = true` limits the exposed tools to `lock_outdated`, `update_plan`, and `lock_diff`.

— names all four tools and the config key, no example given v5.8.0

└──▷ **BREAKING ON UPGRADE**

! `profile = "skillspector"` in `[[lint.external]]` is no longer accepted; the configuration fails to load with 'Value skillspector should be one of the allowed values: agnix, claude-plugin-validate, cisco-skill-scanner, snyk-agent-scan'. Replace it with an explicit entry using `command = ["skillspector", "scan", "--no-llm", "--format", "sarif", "{skill_dirs}"]`, `format = "sarif"`, `inputs = ["skills"]`, `egress = false`, `batch_size = 1`, `version = ">=2, <3"`, `message_properties = ["evidence", "snippet"]`. The profile keys `required_flags`, `offline_notices`, and a profile-level default for `message_properties` are removed.

WAS THIS USEFUL? 👍 👎

Anthropic Claude Code

Feedback ⓘ v2.1.312 NOTES ↗

CODE ↗ RANK

Claude Code is Anthropic's terminal coding agent that plans, edits, and tests code in local repositories.

> Claude Code's v2.1.312 release adds OIDC claim selection for gateway identity, new environment variables to control the system prompt and refusal-fallback behavior, cloud routine/environment UI enhancements, and tighter hook concurrency limits, while discontinuing the `claude_code.safety_event` telemetry event.

└──▷ **WHAT SHIPPED · 9 FEATURES** most completely described first ? what's the number?

01 System prompt control via CLAUDE_CODE_SIMPLE_SYSTEM_PROMPT

85

BREAKING

Adds `CLAUDE_CODE_SIMPLE_SYSTEM_PROMPT=0` env var to keep the full system prompt when using Bedrock, Vertex, Foundry, or Mantle with unrecognized model names such as custom deployment names or inference profile ARNs. These providers now default to a shorter system prompt for unrecognized models; set `CLAUDE_CODE_SIMPLE_SYSTEM_PROMPT=0` to restore the full prompt.

Use Microsoft Entra ID's `oid` claim instead of `sub` to identify users through the Claude apps gateway.

```
$ CLAUDE_CODE_SIMPLE_SYSTEM_PROMPT=0 claude
```

— Exact env var, default-change note, and runnable command v2.1.312

02 Refusal fallback disable via CLAUDE_CODE_DISABLE_REFUSAL_FALLBACK

NEW

80

Adds `CLAUDE_CODE_DISABLE_REFUSAL_FALLBACK=1` env var to turn off automatic model switches after a flagged message, now covering unrecognized custom deployment names in addition to standard models.

Prevent Claude Code from automatically switching models after a flagged message when using custom deployment names on Bedrock or Vertex.

```
$ CLAUDE_CODE_DISABLE_REFUSAL_FALLBACK=1 claude
```

— Named env var with runnable command example v2.1.312

03 **OIDC claim selection for gateway identity** NEW 70

Adds `oidc.user_id_claim` to the Claude apps gateway config, enabling identification by an id_token claim other than `sub` — for example Microsoft Entra ID's `oid`.

— Config key named but no file path or runnable example v2.1.312

04 **Hook concurrency limits for prompt and agent hooks** IMPROVED 70

Changes prompt and agent hooks to run at most 24 concurrently per session (sub-agents included), plus at most 12 past their event timeout; `timeout` values under 1 second are treated as 1 second.

— Exact concurrency limits and timeout floor specified v2.1.312

05 **Subagent autoCompactWindow now lowers compact threshold**

IMPROVED 70

Changes a subagent's `autoCompactWindow` to now lower the window set by `CLAUDE_CODE_AUTO_COMPACT_WINDOW` (previously had no effect); it still cannot raise that window.

— Before/after behavior and exact config names given v2.1.312

06 **Removal of claude_code.safety_event telemetry** DEPRECATED 60

The `claude_code.safety_event` OpenTelemetry event is no longer emitted; pipelines or dashboards consuming it will stop receiving data.

— Exact event name and impact named, no migration path v2.1.312

THINNER COVERAGE BELOW

07 **Bash editor in cloud environment setup script** NEW 55

Adds a Bash editor with colors, line numbers, and a Copy button to the setup script box in cloud environment settings.

— UI path described but limited mechanism detail v2.1.312

08 **Click-to-send for queued messages in fullscreen** NEW 50

Adds click-to-send in fullscreen: clicking the to send now hint under a queued message sends it immediately.

— UI interaction described with exact trigger v2.1.312

09 Effort level selector in cloud routine model picker NEW

45

Adds an effort level selector to the model picker when creating or editing a cloud routine.

— UI addition named but no detail on effort levels v2.1.312

└──▷ BREAKING ON UPGRADE

! Bedrock, Vertex, Foundry, and Mantle now use the shorter system prompt for unrecognized model names by default; set `CLAUDE_CODE_SIMPLE_SYSTEM_PROMPT=0` to restore the full prompt.

! The `claude_code.safety_event` OpenTelemetry event is no longer emitted; pipelines or dashboards consuming it will stop receiving data.

WAS THIS USEFUL? 👍 👎

gentle-ai Feedback ⓘ v4.3.1 NOTES ↗

Gentle-AI configures the AI coding agents you already use: Claude Code, Cursor, OpenCode, Codex, Pi, and more. Choose persistent memory, Organic-Driven Development, curated skills, MCP servers, personas, and optional bounded review.

Adds scoped component updates that rebuild only changed components and carry operator state; launchers `gs` and `gentle-shell`; and acknowledgement advancing the reviewed boundary.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v4.3.1 https://github.com/Gentleman-Programming/gentle-ai.git
# already have the repo? check out this version:
$ git checkout v4.3.1
```

NEW SUPPORTED PLATFORM UPDATED SUPPORTED PLATFORM Adds `gs` and `gentle-shell` launchers published by installs, with `~/.local/bin` added to PATH after launcher publication and `install.sh` / `install.ps1` adding the gentle-ai install directory to PATH.

Adds scoped component updates: installations record component pins and the update check rebuilds only changed components, carrying operator state across rebuilt agents and Windows updates.

Acknowledging an approved review now advances the reviewed boundary so the next review starts from the acknowledged candidate.

NEW INTEGRATION GitHub Copilot CLI agent adapter

└──▷ BREAKING ON UPGRADE

! Run `gentle-ai sync` after upgrading: the Engram protocol and orchestrator guidance changed (high-risk item 1 now covers only data that cannot be rebuilt), and managed agent, reviewer, and runtime assets must match the installed binary.

WAS THIS USEFUL? 👍 👎

DeepSeek Harness

Feedback ⓘ dsh-v0.2.2-alpha.2 CODE ↗

RANK

DeepSeek Harness is an open-source agent harness for building coding agents with plugins.

> DeepSeek Harness shipped an experimental Statistics view for inspecting agent conversation performance, alongside version-compatibility support in the storage subsystem.

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

01 **Statistics tab via client-ui-perf plugin** NEW 78

Adds the `client-ui-perf` conversation view plugin (experimental), surfaced through Developer Tools as the Statistics tab. The performance table gains `pretty-print and wrap` raw JSON inspection, prompt image preview, step start timestamps shown with date and milliseconds, a totals row, context-basis controls, classification of what opened each turn in the question column, non-step event rows, context delta switches, explanations of retries, attempts, context refreshes, and approvals rendered as event rows, and expansion of the full model answer from step details rendered as Markdown.

— Names the plugin and UI path but not an exact command dsh-v0.2.2-alpha.2

THINNER COVERAGE BELOW

02 **Compatible whole-unit versions in storage subsystem** NEW 25

The `storage` subsystem now supports compatible whole-unit versions.

— Mentions the subsystem but gives no mechanism or scope dsh-v0.2.2-alpha.2

WAS THIS USEFUL? 👍 👎

pairlet

Feedback ⓘ v2.6.1 NOTES ↗ RANK

Pairlet (formerly CC Pocket / cc-pocket) — Continue your local AI coding tasks from phone, tablet, or desktop.

> Pairlet v2.6.1 migrates official connections to a new relay domain, adds Codex quota reset cards and mobile-initiated update requests for paired computers, plus chat page UI refinements.

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

01 **Official relay domain migration** IMPROVED 55

Official connections now use `relay.pairlet.cn`; existing paired identities are preserved and legacy connection entries remain compatible.

— Names new domain and compatibility guarantee, but no user action needed v2.6.1

02 **Mobile-initiated paired-computer updates** NEW 55

Mobile settings support requesting updates for compatible paired computers, with progress and restore-status display.

— Clear capability and UI location, no exact navigation path given v2.6.1

03 **Codex quota reset cards** NEW 50

Codex usage detail now surfaces account quota reset cards, requiring explicit confirmation before use.

— Describes behavior and confirmation step, lacks further mechanism detail v2.6.1

04 **Chat page UI refinements** IMPROVED 45

Agent reply copy entry moved to the end of the message body for easier post-read copying, and the chat page adopts a streamlined top bar with usage details now accessed via the quota area and quick menu.

— Two minor UI tweaks grouped, described only in prose v2.6.1

WAS THIS USEFUL? 👍 👎

mvschwarz OpenRig

Feedback ⓘ v0.7.4 NOTES ↗ RANK

OpenRig is a network manager for AI coding agents: build persistent teams with roles, shared context and owned work.

> OpenRig v0.7.4's headline is experimental hosted assistant members that join a rig over HTTP MCP, backed by OAuth discovery and external grant-service key checking, alongside a new conversation-resume lookup and three breaking changes to daemon kernel restore, Slack park resolution, and work-wake routing.

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

01 External grant-service key checking for HTTP MCP NEW

88

`rig mcp serve --http --grant-service <url>` checks each incoming request's key with an external grant service instead of the local token file, with no caching.

Delegate key validation to an external grant service instead of the local token file — useful when tokens are managed centrally across multiple hosted assistants.

```
$ rig mcp serve --http --grant-service
https://grants.example.com/check
```

— Exact flag, example command, and explicit no-caching behaviour v0.7.4

02 Seat history and safe conversation resume NEW

88

`rig seat history --rig <id> --node-id <id>` lists the Claude and Codex conversations a specific node could resume; a client can now resume the exact history a person reviewed, with OpenRig refusing rather than substituting another conversation or starting fresh.

List all Claude and Codex conversations a specific node can resume, then pick the exact one to hand back to a client.

```
$ rig seat history --rig <rig-id> --node-id <node-id>
```

— Runnable command plus explicit refusal-over-substitution behaviour v0.7.4

03 HTTP MCP server for hosted assistants NEW

82

`rig mcp serve --http` exposes the MCP server over HTTP so cloud-hosted assistants outside a local terminal can join a rig as a hosted member under a revocable

bearer-token grant.

Expose your rig's MCP server over HTTP so a cloud-hosted assistant (e.g. Grok) can join as a hosted member under a revocable token grant — put this behind your own TLS proxy.

```
$ rig mcp serve --http
```

— Exact command and example given; grant mechanism described only briefly v0.7.4

---

04 **Daemon kernel restore and --wait-for-kernel reporting** 75

BREAKING

A kernel already down before daemon start is no longer restored on that start; `rig daemon start --wait-for-kernel` for a kernel skipped on purpose now prints the reason and exit path instead of timing out.

— Names the flag and the new exit behaviour precisely v0.7.4

---

05 **OAuth protected-resource discovery for MCP clients** NEW 73

`--oauth-resource` and `--oauth-issuer` flags on `rig mcp serve --http` add OAuth protected-resource discovery (RFC 9728) and standard challenges for MCP clients that sign in with OAuth.

— Flags and RFC named but no worked example v0.7.4

---

06 **qitem target for terminal preview** NEW 70

`qitem:<id>` is now an accepted target for `rig terminal open` and the daemon's terminal preview, showing a read-only membership view of the current local owners of a task.

— Target syntax named but no full worked example v0.7.4

---

07 **Hosted-member roster entries and webhook triggers** NEW 65

Roster entries carry `"hosted": true` so assistants living outside a terminal are listed by `rig_team` with their declared capabilities, caveat, and last MCP contact time; webhook trigger support lets OpenRig ring an assistant's endpoint with a pointer to new work when it arrives.

— Names the field and mechanism but gives no runnable command v0.7.4

THINNER COVERAGE BELOW

08 **Dual MCP protocol revision support** IMPROVED 50

`rig mcp serve` now supports MCP protocol revision 2026-07-28 (no `initialize` handshake) alongside 2025 clients on the same endpoint, over stdio or `--http`.

— Names the spec revision but gives no usage detail v0.7.4

09 **Slack park resolution timing change** BREAKING 50

A park on a registered person's address now resolves only from a Slack reply written after its question, not from earlier replies.

— Behaviour change described but no surface name or command v0.7.4

10 **Work-wake notification routed to person** BREAKING 35

Work waiting for a member nothing can wake is now reported to its person, not to the member.

— Short behaviour note with no mechanism or named surface v0.7.4

└──▷ **BREAKING ON UPGRADE**

! A kernel already down before daemon start is no longer restored on that start; `rig daemon start --wait-for-kernel` for a kernel skipped on purpose now prints the reason and exit path instead of timing out.

! A park on a registered person's address now resolves only from a Slack reply written after its question, not from earlier replies.

! Work waiting for a member nothing can wake is now reported to its person, not to the member.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Code

Feedback ⓘ sdk-typescript-v0.1.24

NOTES ↗ RANK

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

> Qwen Code's two releases focus on managed-session automation — durable message queuing, an external-scratch runtime, session permission rules and `/autofix-pr status` — alongside new settings for environment injection, context window and auto-memory, while renaming `disable*` settings to `enable*`, removing `read_many_files` /Smart Edit/ClearcutLogger, and halving the free OAuth quota.

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

01 Dedicated thread pool for automation scanner rows NEW

75

Adds `automation.scanner-threads` config key to run managed automation scanner rows on a dedicated thread pool, leasing rows from claim time.

Offload managed automation scanner rows to a dedicated thread pool so they don't compete with interactive sessions.

yaml

```yaml
automation:
  scanner-threads: 4
```

— Named config key with mechanism and runnable example v0.26.2

02 Env, context-window and auto-memory config keys NEW

73

Adds `settings.env` to inject environment variables into sessions via the settings file instead of exporting them per shell, `contextWindowSize` to control context size, and `memory.enableManagedAutoMemory` to control managed auto-memory.

Inject environment variables into Qwen Code sessions via the settings file instead of exporting them in every shell.

```json
{
  "settings.env": {
    "OPENAI_API_KEY": "sk-...",
    "OPENAI_BASE_URL": "https://my-proxy.example.com/v1"
  }
}
```

— Names three config keys with a runnable config example `sdk-typescript-v0.1.24`

03 **New /autofix-pr and /verify-pr PR automation commands** NEW

65

Adds `/autofix-pr status` to list all watches in a clone and resume a watch in a later session, and a `/verify-pr` prompt argument hint showing the PR number and supported flags such as `--publish` and `--watch`.

Check the status of all `/autofix-pr` watches in the current clone and pick up a watch dropped in an earlier session.

```
$ /autofix-pr status
```

— Runnable command with clear usage example Named flags but no worked example shown

`v0.26.2`

04 **Experimental unified agent loop** NEW

65

Adds an `experimental.unifiedAgentLoop` setting so ACP prompts share the core agent loop for model rounds.

Enable the experimental unified agent loop so ACP prompts share the core agent loop for model rounds.

```yaml
experimental:
  unifiedAgentLoop: true
```

— Named config flag with a runnable example `v0.26.2`

THINNER COVERAGE BELOW

05 **External-scratch runtime provider** NEW 55

Adds an opt-in `external-scratch` Runtime provider letting an operator-run HTTP service place Session workers while preserving attestation and recovery.

– Explains mechanism but lacks a config example v0.26.2

06 **Image and PDF delivery for managed runtime providers** NEW 55

Enables image and PDF delivery for managed runtime providers with a 768 KiB base64 media limit and a 896 KiB tool-result limit.

– Concrete size limits but no usage example v0.26.2

07 **Durable message queue for Managed Turn** NEW 50

Adds a durable message queue for messages sent during a Managed Turn, supporting queued text and file references that can be delivered, edited, withdrawn, or resumed.

– Describes behaviour well but no API or command given v0.26.2

08 **Operator endpoint for EXTERNAL Workspace storage ownership** NEW 50

Adds an operator endpoint to register immutable, idempotent EXTERNAL Workspace storage ownership with provider identity and tenant alias.

– Names fields but no endpoint path or method given v0.26.2

09 **Git URL source in Web Shell Add-workspace dialog** NEW 50

Adds a Git URL source to the Web Shell Add-workspace dialog so users can clone a repository and open it as a workspace in the browser.

– Names the exact UI dialog as a starting point v0.26.2

10 **ACP subagent tracking fields** NEW 45

Adds `parentToolCallId` and `subagentType` fields for ACP subagent tracking.

– Named fields but no usage example or mechanism detail sdk-typescript-v0.1.24

11 **Agent Skills promoted to stable with symlink support** IMPROVED 45

Agent Skills moves from experimental to stable, and the skill manager adds symlink support.

– States the change but no mechanism or usage path sdk-typescript-v0.1.24

12 **Web Shell managed-agent controls simplified**

BREAKING

45

Web Shell controls for managed-agent are simplified and `cwd` submissions are retired — any workflow that submitted working-directory changes via that mechanism will no longer function.

— States the break but no migration path given v0.26.2

13 **Qwen OAuth free quota halved** BREAKING

40

Qwen OAuth free quota is reduced from 2000 to 1000 requests per day.

— Concrete numbers but no mitigation or alternative noted sdk-typescript-v0.1.24

14 **X-DashScope-Source-Config request header** NEW

40

Adds `X-DashScope-Source-Config` header to matching QwenCloud and Qianwen AI Platform requests to identify site and client.

— Named header but no usage example v0.26.2

15 **Session permission rules and auto-approval controls** NEW

30

Adds session permission rules and auto-approval controls for managed sessions.

— Thin description with no named keys or mechanism v0.26.2

└──▷ **BREAKING ON UPGRADE**

! Settings renamed from `disable*` to `enable*` naming convention (e.g. negative `disable*` settings are now positive `enable*` keys) — any config referencing the old `disable*` names will break.

! The `read_many_files` tool is removed; the functionality is now available only as a `readManyFiles` utility for user @`-commands`.

! The Smart Edit tool and ClearcutLogger are removed.

! Qwen OAuth free quota reduced from 2000 to 1000 requests per day.

! Web Shell controls for managed-agent are simplified and `cwd` submissions are retired — any workflow that submitted working-directory changes via that mechanism will no longer function.

WAS THIS USEFUL?  

palot

Feedback ⓘ v0.21.0 NOTES ↗ RANK

Just another desktop client for OpenCode 2. Manage projects, sessions, parallel agents, requests, and changes on Linux and macOS.

> Palot v0.21.0 and v0.20.0 add native desktop packaging for Linux and macOS alongside checksum and attestation-based download verification, plus a flexible OpenCode setup flow.

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

01 **Download verification via checksums and attestation** NEW 75

Release downloads can be verified using `SHA256SUMS` checksums included with the release, and desktop packages can be verified for integrity with `gh attestation verify` against GitHub build attestations.

— Gives a runnable verification command and checksum file name. v0.21.0 v0.20.0

02 **Native desktop packages for Linux and macOS** NEW 68

Adds Linux x64 desktop packages: `AppImage`, Debian/Ubuntu `DEB`, Fedora `RPM`, and portable `tar.gz`. Adds macOS Apple Silicon desktop packages: `DMG` and `ZIP` (ad-hoc signed; not Developer ID signed or notarized).

— Names exact package formats per platform but no install commands. v0.21.0

THINNER COVERAGE BELOW

03 **OpenCode installation during setup** NEW 35

OpenCode can be supplied via an existing local installation or downloaded as a verified fallback during setup.

— Describes behavior but no setup commands or config keys. v0.21.0

04 **Manual updates for legacy Palot 0.11** BREAKING 35

No auto-update feeds are published for legacy Palot 0.11; updates are manual from this release onward.

— States the change but no migration steps provided. v0.21.0 v0.20.0

└──▷ **BREAKING ON UPGRADE**

! No auto-update feeds are published for legacy Palot 0.11; updates are manual.

! No legacy Palot 0.11 auto-update feeds are published — updates are manual from this release onward.

WAS THIS USEFUL? 👍 👎

All Hands AI OpenHands

Feedback ⓘ v1.34.0 NOTES ↗

RANK

OpenHands is an open-source software-development agent that writes code, runs commands, and browses the web.

> OpenHands v1.34.0 introduces per-conversation Docker runtime configuration via a new Settings → Runtime page, alongside automation tag management and sidebar conversation improvements.

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

01 **Per-conversation Docker runtime via Settings → Runtime** NEW

70

Adds a Settings → Runtime page in the canvas interface that lets users configure Docker runtime options on a per-conversation basis, enabling an isolated agent sandbox for each conversation.

Configure a per-conversation Docker runtime for an isolated agent sandbox.

> 📍 In the canvas, go to Settings → Runtime and set the Docker runtime options for the active conversation.

– Names exact UI path and includes runnable steps v1.34.0

THINNER COVERAGE BELOW

02 **Start new conversation in sidebar sandbox** NEW 40

Enables starting a new conversation directly in the row's sandbox from the sidebar conversation list.

– Describes action and location but no exact UI path v1.34.0

03 **Filter and edit controls for automation tags** NEW 30

Adds filter and edit controls for automation tags in the automations list.

– Names the area but no mechanism or UI path v1.34.0

04 **Org-scoped share toggle label** IMPROVED 25

Labels the share toggle Share with Org users in team org contexts to clarify sharing scope.

– Minor copy change, no further mechanism given v1.34.0

WAS THIS USEFUL?  

agents-cli

Feedback ⓘ v1.12.0 NOTES ↗ RANK

The CLI and skills that turn any coding assistant into an expert at creating, evaluating, and deploying AI agents on Google Cloud.

> agents-cli v1.12.0 adds a Claude Agent SDK extension template, promotes ADK Go to General Availability, renames the Python template to `adk_python`, and ships several deploy, scaffold-enhance, and info improvements alongside a few breaking changes.

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

01 Claude Agent SDK extension scaffold template NEW 90

A new `agents-cli scaffold create --agent google/agents-cli/extensions/claude/template@<version>` bootstraps a Claude Agent SDK project that deploys to Agent Runtime or Cloud Run over A2A, with Cloud Trace, Cloud Logging, Cloud Monitoring, optional BigQuery Agent Analytics, and Cloud SQL sessions.

Bootstrap a new Claude-based agent project that deploys over A2A with built-in observability and optional BigQuery analytics.

```
$ agents-cli scaffold create --agent google/agents-cli/extensions/claude/template@v1.12.0 my-claude-agent
```

— Names the exact command and every built-in integration v1.12.0

02 scaffold enhance works on non-agents-cli projects NEW 83

`agents-cli scaffold enhance <template>` now works on projects agents-cli did not create: it writes the template's files as `create` would and moves replaced files (such as `pyproject.toml`, `README.md`, `Dockerfile`, `.env`) to `.agents-cli-original/`, keeping their original paths.

Migrate an existing project to the Claude Agent SDK extension template, with replaced files safely backed up to `.agents-cli-original/`.

```
$ agents-cli scaffold enhance google/agents-cli/extensions/claude/template@v1.12.0
```

— Names command, backup mechanism, and affected files v1.12.0

03 scaffold enhance --dry-run exit code change

BREAKING

78

`agents-cli scaffold enhance --dry-run` now exits with code 2 (was 0) when combined with `--force`, when run with `-y` and no override flags, or when run without a saved `agents-cli-manifest.yaml`; scripts checking for exit code 0 need updating.

— Exact flag combinations and new exit code given `v1.12.0`

04 **Cloud Run Terraform provider pin bump** BREAKING

78

Cloud Run projects now use the `hashicorp/google-beta` provider pinned to `~> 8.7`; after running `agents-cli scaffold upgrade`, if `terraform init` reports a locked `google-beta` older than 8.7, run `terraform init -upgrade` once.

— Exact provider version and remediation command given `v1.12.0`

05 **ADK Go template promoted to GA** IMPROVED

75

The `adk_go` template is now General Availability; `agents-cli create --agent adk_go` accepts ADK Go v2 templates whose `manifest.yaml` declares `language: go`, and rejects templates built for other ADK Go versions.

Create a new ADK Go v2 agent project now that `adk_go` is GA and production-ready.

```
$ agents-cli create my-go-agent --agent adk_go -y
```

— Clear GA status, command, and manifest validation rule `v1.12.0`

06 **Python template renamed to adk_python** BREAKING

73

The Python agent template is renamed to `adk_python`, selectable with `--agent adk_python` or `-a adk_python`; `adk` remains a valid alias, and new projects record `adk_python` in `agents-cli-manifest.yaml`. Projects that record `adk_python` in that manifest require this CLI version or newer to run `enhance` or `upgrade`.

— Names flags, alias, manifest field, and upgrade requirement `v1.12.0`

07 **agents-cli deploy --dry-run validation improvements** IMPROVED

65

`agents-cli deploy --dry-run` for Agent Runtime now prints the plan even when a check fails, and warns (rather than silently skips) when the Secret Manager API is disabled; it also now validates Cloud Run deployments against Google Cloud.

— Names flag and three concrete behavior changes `v1.12.0`

08 **Configurable session types for extension templates** NEW

62

Extension templates can set `settings.session_types` in their template manifest, so `create` offers and accepts only those session types, with the first listed as the default.

— Names manifest key and behavior but no example `v1.12.0`

THINNER COVERAGE BELOW

09 **Nested agent template directories** NEW

45

Supports nested agent template directories (e.g. `src/my_agent`) in both remote templates and with `--agent-directory`.

— Names flag but mechanism is thin `v1.12.0`

10 **Globally installed skills listing in agents-cli info** NEW

35

`agents-cli info` now lists globally installed skills.

— Bare description with no further detail `v1.12.0`

└──▷ **BREAKING ON UPGRADE**

! `agents-cli scaffold enhance --dry-run` now exits with code 2 instead of 0 when combined with `--force`, when run with `-y` and no override flags, or when run without a saved `agents-cli-manifest.yaml` — scripts that checked for exit code 0 will need updating.

! New and upgraded projects require this CLI version or newer to run `enhance` or `upgrade` on projects that record `adk_python` in `agents-cli-manifest.yaml`.

! Cloud Run projects now use the `hashicorp/google-beta` provider pinned to `~> 8.7`; after running `agents-cli scaffold upgrade`, if `terraform init` reports a locked `google-beta` older than 8.7, run `terraform init -upgrade` once.

WAS THIS USEFUL? 👍 👎

repoprompt-ce

Feedback ⓘ v1.7.0 NOTES ↗ RANK

Repoprompt-ce is a native macOS context engineering app for AI coding agents, with an MCP CLI (community edition of RepoPrompt).

> RepoPrompt CE's v1.7.0 release adds re-prompt drafting, remappable keyboard shortcuts, nested Overseer delegation scopes, and new GPT-6.1/Codex 0.164.0 model choices, alongside improvements to Oracle attachments and the Model Router.

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

01 Remappable keyboard shortcuts for commands and composer NEW

80

Adds remappable keyboard shortcuts for command and composer actions, including `stop-run`, `focus-composer`, `reopen-stashed-session`, and `jump-to-attention-session` bindings, configurable via Settings → Keyboard Shortcuts.

Remap keyboard shortcuts to fit your workflow — for example, bind a key to immediately stop a running agent or jump to a session needing attention.

📍 In the application, go to Settings → Keyboard Shortcuts, locate the command or composer shortcut you want to change (e.g., stop-run, focus-composer, reopen-stashed-session, jump-to-attention-session), and assign your preferred key binding.

— Names exact bindings and exact settings path to configure them v1.7.0

02 Compatible local Codex executable path for Computer Use NEW

80

Adds a compatible local Codex executable path setting under Settings → Codex CLI → Advanced, letting users point RepoPrompt CE at an OpenAI-signed Codex executable instead of the bundled runtime to enable Computer Use workflows; requires restarting the app after setting it.

Enable Computer Use by pointing RepoPrompt CE at an OpenAI-signed Codex executable instead of the bundled runtime.

📍 In the application, go to Settings → Codex CLI → Advanced, set the path to your compatible local Codex executable, then restart RepoPrompt CE.

— Exact settings path plus runnable steps including restart v1.7.0

03 Overseer nested delegation scopes and worktree operations

NEW

60

Adds nested delegation scopes and worktree operations to Overseer agents, enabling Overseers to manage sessions and operate worktrees on behalf of linked agents.

— Describes new agent mechanism but no setup steps given v1.7.0

THINNER COVERAGE BELOW

04 **Re-prompt capability for drafts** NEW

50

Adds re-prompt capability: rewrite a draft inline, review clarifications, and select the model and guidance used before sending.

— Explains the flow but no exact UI path or command v1.7.0

05 **Model Router and Auto Effort expansion** IMPROVED

40

Expands Model Router and Auto Effort with additional decision services, stronger defaults for new tasks, and improved usage balancing.

— Vague on which decision services or defaults changed v1.7.0

06 **Oracle attachment and results improvements** IMPROVED

40

Adds `oracle_send` image support, expanding Oracle workflows to include image attachments alongside file attachments; also expands general file attachment support and adds clearer Oracle group results and guidance display.

— Names oracle_send but other additions described only generically v1.7.0

WAS THIS USEFUL? 👍 👎

Sourcegraph Amp

Feedback ⓘ Amp on Android (2026-10-27)

BLOG ↗ RANK

Amp is Sourcegraph's agentic coding tool for the terminal and editor, running multi-step edits with subagents and shared team threads.

> Amp shipped native companion apps for macOS, iOS, and Android that let users follow agent sessions remotely, adding notifications, dictation, and cross-app sharing.

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

01 **macOS, iOS, and Android companion apps** NEW 56

Amp added native apps for macOS, iOS, iPad, and Android to follow agents running elsewhere, including a native Android app downloadable from `ampcode.com/app`. The apps add notifications support on macOS and iOS, plus dictation and sharing content in from other apps on mobile.

— Names platforms and a download URL but no deeper mechanism. Amp on Android (2026-10-27)

WAS THIS USEFUL? 👍 👎

SST OpenCode

Feedback ⓘ v2.2.0 NOTES ↗ RANK

OpenCode is an open-source AI coding agent for terminal-based development tasks.

> OpenCode v2.2.0 adds a keyed session-instructions API for plugins, experimental Desktop plugin support and image/decision generation APIs, alongside several breaking changes to package structure and dependencies.

└──▷ **WHAT SHIPPED · 7 FEATURES** most completely described first ? what's the number?

01 Keyed session instructions API for plugins NEW 95

Plugins can call `ctx.session.instructions.entry.put`, `ctx.session.instructions.entry.list`, and `ctx.session.instructions.entry.remove` to attach and update keyed session instructions without editing the system prompt or busting the prompt cache. Entries survive compaction and reverts and do not trigger an extra model turn.

Keep persistent per-session context in a plugin without invalidating the prompt cache on every update.

```text
ctx.session.instructions.entry.put('my-plugin-state', 'User
prefers concise answers in TypeScript.');
// Later, update without triggering a new model turn:
ctx.session.instructions.entry.put('my-plugin-state', 'User
prefers verbose explanations in Python.');
// Remove when no longer needed:
ctx.session.instructions.entry.remove('my-plugin-state');
```

— Named methods plus a runnable usage example v2.2.0

02 OpenTUI dependency bump to 0.7.0 BREAKING 90

`@opencode/plugin` now requires `@opentui/core` and `@opentui/solid` 0.7.0 (up from 0.6.8); plugins depending on OpenTUI 0.6 will fail to install until upgraded. Migrating to OpenTUI 0.7 requires assigning `renderBefore`, `renderAfter`, and `onLifecyclePass` in the constructor instead of as class fields, calling refreshHooks() after adding a hook by reflection, and replacing defineNativeIntegration() with a static `nativeIntegration` declaration.

— Exact versions and concrete migration steps given v2.2.0

03 @opencode/ai package restructuring BREAKING 80

The `@opencode/ai` Promise API moves to the package root to match `@opencode/sdk` and `@opencode/client`; the Effect API relocates to `@opencode/ai/effect`; `@opencode/ai/promise` is removed; and the wildcard subpath export is replaced with an explicit list of provider and protocol entry points. Code relying on the old `@opencode/ai/promise` import or wildcard imports must be updated.

— Named paths but no migration command shown v2.2.0

THINNER COVERAGE BELOW

04 **Experimental image and decision generation APIs** NEW

40

Adds experimental image generation and decision generation APIs to plugins and the SDK, enabling image generation and decision model workflows (such as typesafe, clef).

— Example workflows named but no API surface specified v2.2.0

05 **GitHub Action no longer published** BREAKING

35

The OpenCode GitHub Action is no longer published; remove it from your workflows.

— Clear instruction but no further detail v2.2.0

06 **Desktop plugin create, install, and share support** NEW

30

Adds create, install, and share support for Desktop plugins (experimental).

— No commands or UI path described v2.2.0

07 **Configurable prompt cache TTLs** NEW

25

Adds configurable prompt cache TTLs per provider/model.

— No config key or mechanism given v2.2.0

└──▷ BREAKING ON UPGRADE

! `@opencode/ai/promise` is removed; update imports to use the package root for the Promise API instead.

! The wildcard subpath export on `@opencode/ai` is replaced with an explicit list of provider and protocol entry points; any code relying on wildcard imports must be updated.

! `@opencode/plugin` now requires `@opentui/core` and `@opentui/solid` 0.7.0 (up from 0.6.8); plugins depending on OpenTUI 0.6 will fail to install until upgraded. In OpenTUI 0.7: assign `renderBefore`, `renderAfter`, and `onLifecyclePass` in the constructor instead of as class fields; call refreshHooks() after adding a hook by reflection; replace defineNativeIntegration() with a static `nativeIntegration` declaration.

! The OpenCode GitHub Action is no longer published; remove it from your workflows.

WAS THIS USEFUL?  

◆ Model Training & Fine-Tuning

fizgig

Feedback ⓘ v8.9.93 NOTES ↗ RANK

Fizgig — LoRA & Fine-tune Studio for Klein 9B, Krea 2, MiniMax H3, Qwen Image 2.1, Anima & Z-Image Turbo and Ace-Step XL: train, fine-tune, profile, repair and extract LoRAs & LoKR for Image, Video & Music Models

> Fizgig v8.9.93 adds pod version-locking via a new environment variable, eight selectable UI themes including accessibility options, and training-tab controls to exclude specific prompts from slider adjustments.

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

01 **Pod version-locking with FIZGIG_REF** NEW 72

Adds the `FIZGIG_REF` environment variable to pin a pod to a specific Fizgig release, overriding the automatic version-matching that normally syncs the pod to the host computer's version.

— Named env var with clear override behaviour, no setup steps shown v8.9.93

02 **Eight new UI themes** NEW 68

Adds eight UI themes selectable under Preferences → Theme : Fizgig , Light n' Orange , Seventies , Synthwave , Graphite , Colour-blind safe , High contrast , and Dim , each stored per-device including accessibility-focused options.

— Names all eight themes and exact UI path v8.9.93

THINNER COVERAGE BELOW

03 **Prompt exclusion from slider dials** NEW 58

Adds Leave these alone and How firmly fields to prompt sliders on the Training tab so nominated prompts are excluded from dial changes during training.

— Names fields and location but no mechanism detail v8.9.93

04 **MOSS-Image default preset changed to Fast** IMPROVED 45

Changes the MOSS-Image default preset to Fast (rank 16, 0.25 MP).

— Concrete values given but no further context on impact v8.9.93

05 **Rotating tips on loading screen** NEW 15

Adds a rotating tip display to the Fizgig loading screen.

— Bare description, minor cosmetic addition v8.9.93

WAS THIS USEFUL? 👍 👎

Axolotl AI Axolotl

Feedback ⓘ v0.21.1 NOTES ↗ RANK

Axolotl fine-tunes open language models from YAML configuration, supporting full fine-tuning, LoRA, QLoRA, and multi-GPU training recipes.

> Axolotl v0.21.1 adds background checkpoint mirroring and a rank-0-gather-free sharded LoRA/QLoRA checkpoint format, plus per-group eval metrics, merge-aware BF16 LoRA for RL, multi-node vLLM serving, and several SonicMoE and NVFP4 performance-tuning controls.

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

01 **Sharded FSDP2 LoRA checkpoint format** BREAKING 95

Adds `adapter_checkpoint_format: sharded` to checkpoint FSDP2 LoRA and QLoRA runs without a rank-0 gather, cutting a Beam `EP=8` checkpoint stall from ~2:20 to 11.6 s, with examples in `examples/expert_parallel/`. Checkpoints saved model-only or sharded this way can no longer be resumed by older Axolotl versions.

Eliminate the rank-0 gather stall when checkpointing large EP LoRA runs, reducing a Beam `EP=8` checkpoint from ~2 min 20 s to 11.6 s.

yaml

```yaml
adapter_checkpoint_format: sharded
```

— Exact benchmark numbers, config key, and breaking note v0.21.1

02 **Background checkpoint mirroring to shared storage** NEW 77

Adds `checkpoint_mirror` and `checkpoint_save_retry` config keys so Axolotl saves checkpoints to fast local storage while copying them to shared storage in the background, automatically pausing and retrying instead of crashing when the disk fills up.

Mirror checkpoints to shared storage in the background during a long LoRA run so training is never blocked by slow network writes, and stall instead of crashing on a full disk.

yaml

```yaml
checkpoint_mirror: gs://my-bucket/checkpoints
checkpoint_save_retry: true
```

— Mechanism and exact config keys given, no numeric benchmark v0.21.1

03 **Merge-aware BF16 LoRA training for RL** NEW 75

Adds `lora_merge_aware: bf16` to train LoRA RL on exactly the BF16-rounded weights that a merged vLLM sync serves, plus `trl.vllm_merge_check_interval` to log how much of each LoRA update survives rounding.

— Named config keys and purpose, no numeric example v0.21.1

04 **Multi-node vLLM serving** NEW 75

Adds multi-node support for `axolotl vllm-serve` via `vllm.nnodes`, `vllm.master_addr`, and `--node-rank` so one vLLM server can span several nodes.

— Exact flags named, no scaling limits given v0.21.1

05 **Compiled NVFP4 LoRA merge path** NEW 75

Adds opt-in `merge_compile` for NVFP4 LoRA merges, with batched pinned transfers; a rank-128 merge into Beam is at least 27% faster with 8 GB less peak memory.

— Concrete speed and memory numbers, config flag named v0.21.1

06 **Per-dataset-group eval metrics** NEW 70

Adds `eval_name` field to `test_datasets` entries for per-group eval metrics (e.g. `eval/math/loss`) alongside the aggregate loss.

Track loss separately for each evaluation domain (e.g. math vs. code) so you can detect per-task regressions during training.

yaml

```yaml
test_datasets:
  - path: my_org/math_eval
    type: chat_template
    eval_name: math
  - path: my_org/code_eval
    type: chat_template
    eval_name: code
```

— Config key and example given, limited further mechanism v0.21.1

07 **SonicMoE autotune and LoRA split controls** NEW 70

Adds `quack_autotune: false` to skip SonicMoE autotune candidate sweeps, `autotune_profiles` to pin quack configs (new row-count results now reuse a power-of-two bucket's winner by default), and experimental `sonicmoe_lora_split` to run expert LoRA as separate low-rank GEMMs.

— All config keys named, each individually thin v0.21.1

08 **Full assistant-turn training for Harmony templates** IMPROVED

65

Adds `generation_tags: true` on a `chat_template` dataset to train the full assistant turn on Harmony-style templates, including channel headers that were previously left untrained.

— Names config flag and before/after, no runnable example v0.21.1

09 **activation_offloading legacy option changes** BREAKING

65

`activation_offloading: legacy` is now a deprecated alias of `true`; `activation_offloading: disk` (with `gradient_checkpointing: offload_disk`) is now rejected outright.

— Exact config values and rejection behavior named v0.21.1

10 **CoreWeave CKS node-local model loading** NEW 60

Adds CoreWeave CKS `host_paths` config key to load `base_model` from node-local NVMe instead of shared PVC.

— Names config key and storage target, no example v0.21.1

THINNER COVERAGE BELOW

11 **Random label dropping during SFT** NEW 45

Adds `label_keep_prob` config key to randomly drop labels of chosen token IDs during SFT.

— Bare config key description with little mechanism v0.21.1

└──▷ **BREAKING ON UPGRADE**

! EP checkpoints saved model-only or sharded with `adapter_checkpoint_format: sharded` can no longer be resumed by older Axolotl versions.

! `activation_offloading: legacy` is now a deprecated alias of `true`; `activation_offloading: disk` (with `gradient_checkpointing: offload_disk`) is now rejected outright.

WAS THIS USEFUL?  

AGENT

◆ AI Agent Frameworks

Arcade

Feedback ⓘ snapshot-20261028 NOTES ↗ RANK

Arcade provides an actions runtime that lets AI agents securely call external services with authorization and governance controls.

> Arcade's snapshot-20261028 release centers on secret governance and deployment hardening — new per-scope secret withholding for API-registered workers, SMTP invitation delivery, cloud-credential passthrough to deployed servers, redaction of secrets in tool responses and logs, and a stricter server readiness check that can break non-conforming deployments.

└──▷ **WHAT SHIPPED · 9 FEATURES** most completely described first ? what's the number?

01 Secret withholding for API-registered workers NEW

86

Adds `tools.worker_secret_sharing` (Helm: `engine.workerSecretSharing`) to withhold secrets from API-registered workers, with scoping controls per `static`, `organization`, and `project` level.

Prevent API-registered workers from receiving org-scoped secrets while still allowing static workers to access them.

yaml

```yaml
tools:
  worker_secret_sharing:
    static: true
    organization: false
    project: false
```

— Exact config and Helm keys with a working example config. snapshot-20261028

02 ServiceAccount token mount for Kubernetes deployments NEW

81

Adds `engine.serviceAccount.automountToken` (default `false`) to mount the engine's ServiceAccount token for deployments using `deployments.kubernetes` configured via `engine.extraConfig`.

Mount the engine ServiceAccount token so Kubernetes-native deployments can authenticate to the cluster API via `engine.extraConfig`.

```yaml
engine:
  serviceAccount:
    automountToken: true
  extraConfig:
    deployments:
      kubernetes: {}
```

— Named config key, default value, and example given. snapshot-20261028

03 **SMTP support for coordinator invitation emails** NEW

78

The coordinator can now send invitation emails through SMTP, using new `coordinator.email.*` config to route them through a custom SMTP server instead of only SendGrid.

Route coordinator invitation emails through your own SMTP server instead of SendGrid.

```yaml
coordinator:
  email:
    host: smtp.example.com
    port: 587
    username: arcade-invites@example.com
    password: <smtp-password>
```

— Named config key plus a runnable SMTP example. snapshot-20261028

04 **Stricter server readiness check requirements** BREAKING

70

A redeployed server must not bind `0.0.0.0` and must serve `/worker/health` without auth — existing server images that do not satisfy these requirements will fail the new stricter readiness check.

— Exact requirements and migration impact named. snapshot-20261028

05 **MCP server runner image dependency installation** IMPROVED

62

The MCP server runner image now includes `uv` and `git`, installing declared Python dependencies from `uv.lock`, `requirements.txt`, or `pyproject.toml`.

— Names tools and files but no usage walkthrough. snapshot-20261028

THINNER COVERAGE BELOW

06 **Cloud credentials delivered to deployed servers** IMPROVED

57

Deployed servers that declare `AWS_*`, `GOOGLE_*`, or `AZURE_*` env vars/secrets now actually receive those cloud-provider credentials.

– Names env var patterns but gives no setup example. snapshot-20261028

07 **Secret redaction in tool responses and deployment logs**

IMPROVED

56

Tool call responses now redact exact copies of the credentials supplied to that call before they reach the caller, logs, or hooks, and deployment logs also redact secrets at write time. Log redaction only applies once a deployment runs on the new runner image — logs from older runner images are unaffected.

– Mechanism and scope clear but no config surface to act on. snapshot-20261028

08 **Clear saved OAuth Client ID from dashboard** NEW

50

Dashboard lets an operator clear a remote MCP server's saved OAuth Client ID, also dropping its paired secret.

– UI action named without exact navigation path. snapshot-20261028

09 **In-cluster usage-reporting collector** NEW

45

Adds an opt-in in-cluster usage-reporting collector so deployments no longer require an outbound reporting connection.

– No config key or activation steps named. snapshot-20261028

└──▷ **BREAKING ON UPGRADE**

! A redeployed server must not bind `0.0.0.0` and must serve `/worker/health` without auth — existing server images that do not satisfy these requirements will fail the new stricter readiness check.

! Deployment-log secret redaction only takes effect once the deployment runs on the new runner image; logs from older runner images are unaffected.

WAS THIS USEFUL?  

cotal

Feedback ⓘ v0.115.4 NOTES ↗ RANK

The open standard for agent coordination

> Cotal's biggest window centers on multi-space hosting: one `deliver` or `auth-service` process can now serve many spaces from a single manifest, with partial-start resilience when spaces fail to open, per-space service-endpoint control, and bounded, auto-expiring issued-authority stores plus a migration command for pre-0.111.0 meshes. Smaller releases hardened `cotal update`, added automatic manager resolution for `stop` / `attach` / `input`, a CLI help Examples section, a `team remove --dry-run` safety net, faster xkey crypto, and Claude Code launch/permission-mode fixes.

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

01 Multi-space hosting manifests for `deliver` and `auth-service`

NEW 98

`cotal deliver --spaces <manifest> --server <url>` (v0.110.0) lets one delivery process hold a lane per space listed in a manifest, rereading it on disk change, `SIGHUP`, or a backstop interval; `cotal auth-service --spaces <manifest> --server <url>` (v0.112.0) extends this to serve multiple spaces as full primaries from one process, each with its own callout, authority plane, loopback exchange, JWKS port, and `auth-service.json` record, via a version-2 manifest format (`{"version": 2, "spaces": [{"space": "<s>", "exchangePublicPort": 7451, "exchangeTrustedProxy": true}]}`) with per-entry fields `exchangePublicPort`, `exchangePublicUrl`, `exchangeTrustedProxy`, `advertisedServer`, and `agentProvisioningUrl`. `parseSpaceManifest` and `runSpaceHost` are exported for programmatic use, every multi-space host writes `<manifest>.applied` holding the SHA-256 of the applied manifest, and each space's lane opens independently with a 30 second deadline so a hung open can't block others; `--creds`, `--dev-mint`, `--shard`, `--shards`, and `deliver pending` refuse when combined with `--spaces`, `--spaces` requires `--server`, and `--space` / `--spaces` are mutually exclusive. Version-2 manifest entries also gain a `"serviceEndpoints": true|false` field giving per-space control over service endpoint hosts, discovery/callout connections, admission records, and the `/service-endpoint-authority` and `/service-endpoint-host` routes — a missing or `false` value makes both routes return 404 for that lane, and changing the field closes and reopens only that space's lane. A new parseAuthHostManifest(text) function and `AuthHostEntry` type are exported from `@cotal-ai/auth` so programs writing the manifest can validate it with the host's own parser before renaming the file into place. `--service-endpoints` now refuses to be combined with `--spaces` (previously it enabled service endpoints for every lane); set `"serviceEndpoints": true` per space instead.

Run a single delivery process that holds one lane per space listed in a manifest file, replacing one-process-per-space deployments.

```
$ cotal deliver --spaces /etc/cotal/spaces.manifest --server
<url>
```

Host auth service for two spaces from one process using a version-2 manifest, so one space's failure does not take down the other.

```
$ cotal auth-service --spaces ./spaces-manifest.json --server
https://nats.example.com
```

Define a version-2 space manifest assigning each space its own public exchange port before passing it to `--spaces`.

```json
{"version": 2, "spaces": [{"space": "tenant-a",
"exchangePublicPort": 7451, "exchangeTrustedProxy": true},
{"space": "tenant-b", "exchangePublicPort": 7452,
"exchangeTrustedProxy": false}]}
```

Selectively enable service endpoints for only specific spaces in a multi-space auth-service deployment.

```
$ cotal auth-service --spaces manifest.json
```

Per-space manifest showing one space with service endpoints enabled and another with them disabled.

```json
[
  {
    "space": "prod-space",
    "serviceEndpoints": true
  },
  {
    "space": "staging-space",
    "serviceEndpoints": false
  }
]
```

Validate a manifest file with the same parser cotal uses before atomically replacing the live copy.

```javascript
import { parseAuthHostManifest } from '@cotal-ai/auth';

const text = fs.readFileSync('manifest.json', 'utf8');
const result = parseAuthHostManifest(text); // throws with host message on invalid input
fs.renameSync('manifest.json', 'manifest.live.json');
```

— Full commands, manifest schema and library exports all given. Field, routes, function and migration path all named.

v0.115.4 v0.112.0 v0.110.0 v0.114.0

02

Bounded issued-authority stores via per-message TTL

BREAKING

94

`KV_cotal_issued_<space>` and `KV_cotal_accepted_<space>` stores gain `allow_msg_ttl`, a NATS per-message TTL that removes rows 60 seconds after their generation's horizon, encoded in the row's first 8 hex digits, giving every issued and accepted row an automatic expiry tied to bearer, instrument, assignment, or seat lifetime. The `issuer` credential is granted `STREAM.INFO` on the accepted store, and every issuer/resolver session now verifies both stores' shape at bind time. Stores created before v0.111.0 are refused by every binder: stop the space's auth service, callout replicas, and managers, back up both streams with `nats stream backup`, remove them with `nats stream rm`, then start v0.111.0 to recreate them — generations issued before the upgrade are refused afterwards. Cotal also surfaces `cotal up --max-file-store 67108865` as the exact runnable command in the Start here row when a stack's recorded store cap is below the floor.

Migrate existing issued-authority stores before starting v0.111.0 — back up then remove the old streams so the new release can recreate them with `allow_msg_ttl` and horizon-aware TTLs.

```
$ nats stream backup KV_cotal_issued_<space> ./backup/issued &&
nats stream backup KV_cotal_accepted_<space> ./backup/accepted &&
nats stream rm KV_cotal_issued_<space> && nats stream rm
KV_cotal_accepted_<space>
```

Bring up a new stack when the recorded store cap is below the required floor, using the exact flag Cotal now surfaces in the Start here row.

```
$ cotal up --max-file-store 67108865
```

— Exact store names, TTL mechanism and migration commands given. v0.111.0

03 **`cotal up --recreate-issued-stores` migration command** NEW

90

`cotal up --recreate-issued-stores <backup-dir>` backs up, deletes, and recreates `KV_cotal_issued_<space>` and `KV_cotal_accepted_<space>` stores created before bounded retention, letting meshes on 0.110.x or earlier start on 0.111.0+. It uses a new one-shot `issued-store-recreate` credential profile with `STREAM.INFO`, `STREAM.SNAPSHOT`, and `STREAM.DELETE` permissions, valid for 30 minutes. `cotal up` on a pre-retention store now refuses `CONFLICT` (exit 2) and throws `IssuedStoreBeforeRetentionError` instead of `OPERATION_UNKNOWN` (exit 3), naming `--recreate-issued-stores` as the remedy.

Migrate a mesh that was running 0.110.x or earlier so it can start on 0.111.0+, backing up the pre-retention issued-authority KV stores to a safe directory before they are recreated.

```
$ cotal up --recreate-issued-stores ./issued-stores-backup
```

— Exact command, credential scopes and error codes named. v0.113.0

04 **Automatic manager resolution for stop/attach/input** NEW

90

`cotal stop <name>`, `cotal attach <name>`, and `cotal input <name>` on user-auth meshes with multiple live managers now auto-route to the manager holding the named agent instead of prompting interactively, searching all managers and running on the one that holds it. If no manager holds the name, it returns `TARGET_NOT_FOUND` with the list of managers searched; if some didn't answer, the terminal prompt narrows to only the responding ones. With `--json` or no terminal, refusals include the managers list and a `Next` suggestion that re-issues the command with `--on` pinned to each candidate manager; `--on` can also be used directly to pin a manager and skip the search.

Pin to a specific manager to skip the automatic search when you already know where the agent lives.

```
$ cotal stop my-agent --on manager-prod-1
```

Get a machine-readable refusal with candidate managers and retry commands when running in CI with no terminal.

```
$ cotal stop my-agent --json
```

— Exact flags, error code and JSON envelope behaviour described. v0.109.21

05 **`cotal team remove --dry-run` for stranded deployments** NEW

82

`cotal team remove <id> --dry-run` previews unwinding deployments stranded by an incomplete `team deploy` where the agent exited on launch, retiring each allocated agent via its identity and lifecycle UID and deleting each card at its accepted create revision. A blocked step is now identified by name in the `LIFECYCLE_BLOCKED` error instead of leaving the cause anonymous.

Preview what would be cleaned up before committing to removing a stranded incomplete deployment.

```
$ cotal team remove <id> --dry-run
```

— Exact flag and error identifier named with runnable command. v0.109.16

06 **Partial-start for multi-space hosts when spaces fail to open**

81

BREAKING

`cotal deliver --spaces` and `cotal auth-service --callout-only --spaces <manifest>` now start with whichever spaces they can serve instead of refusing when any single space fails to open; each names the spaces it cannot serve with a reason and retries them every 5 seconds, and `--callout-only --spaces` also reports counts of answering vs. named spaces in its ready line. This changes exit-code semantics: a supervisor that relied on either command exiting nonzero when one space couldn't open now sees the host start instead, and must read the ready line or per-space lines to find spaces still retrying.

Start a multi-space callout host that comes up immediately with whichever spaces are available, rather than failing hard when one space is unreachable.

```
$ cotal auth-service --callout-only --spaces ./manifest.json
```

Before adding a space to a callout host manifest, check whether its service endpoint admission is open so you can avoid spaces that will be perpetually retried.

```javascript
const { admitsServiceEndpoints } = require('@cotal-ai/auth');
const ok = await admitsServiceEndpoints({ space: 'my-space',
server: serverConfig, secrets: secretsConfig });
if (!ok) console.warn('Space refused for admission — exclude from
callout manifest');
```

— Retry interval and exit-code change named; no runnable deliver example. v0.115.0 v0.112.0

07

`LaunchSpec.confirmBlockers` and honest `cotal up -f` reporting

NEW

80

Launch manifests gain an optional `LaunchSpec.confirmBlockers` field to declare first-run blocker screens (such as Claude Code's login method choice, text style picker, or Bypass Permissions warning), so the pty, custody, tmux, cmux, Orca, and Herdr runtimes can name the blocking screen and the step that clears it when a managed seat stalls there. `cotal up -f` now reports that agent launch is pending rather than claiming success, naming `cotal ps --space <space>` and the manager log as where to check for a failed launch and its reason.

Check agent launch status after `cotal up -f`, since the command no longer waits for agents to join or claims they launched successfully.

```
$ cotal ps --space <space>
```

— Field and runtimes named; checking command given. v0.109.14

08

`cotal update` recovery messaging, extension rebind, and config permission hardening

IMPROVED

72

`cotal update` refusals now print the recovery path instead of leaving the next step to guess (e.g. `cotal down --space <space> && cotal update --self` when blocked by a legacy manager), and it now verifies a first-party operator extension already at the running CLI version and rebinds its shared peers instead of running a full npm reinstall. It also removes group and other write bits from files and folders the user owns under `$XDG_CONFIG_HOME/cotal` or `~/.config/cotal`, closing a security gap from installs under umask 0002, and prints `✓ removed group and other write from <n> entries in Cotal's config at <path>`.

When a `cotal update` is blocked by a legacy manager, use the now-printed recovery path to unblock it without guessing the next step.

```
$ cotal down --space <space> && cotal update --self
```

— Recovery command given; extension-rebind mechanism stated but not detailed. Exact paths and printed message named, no example command.

v0.115.4 v0.112.0 v0.109.18

---

09 **Foreground `cotal spawn` of Claude Code preserves permission mode**

64

BREAKING

Foreground `cotal spawn` of Claude Code no longer passes `--permission-mode bypassPermissions`, preserving Claude's own permission mode in the calling terminal. Users who relied on bypass being set automatically must accept the Bypass Permissions prompt manually or pass `--opt permission-mode=default`.

— Exact flags named; no example command shown. v0.109.14

---

10 **Faster xkey shared-key derivation in auth callout** IMPROVED

62

Improves xkey shared-key derivation in the auth callout to once per server key (capped at 64 LRU keys) instead of twice per connect, cutting the dominant X25519 crypto CPU cost; a forged Nats-Server-Xkey header cannot grow the cache. `@cotal-ai/auth` now depends directly on `tweetnacl`, using the same version already installed by `@nats-io/nkeys`.

— Mechanism and cache cap named, no reader-facing command. v0.112.0

THINNER COVERAGE BELOW

---

11 **CLI help Examples section with note annotations** NEW

42

Cotal CLI help now includes an Examples section with working commands for every command and subcommand, and supports a `{ command, note }` object shape for a command's `examples` entry (in addition to bare strings), where the note appears beside or below the command depending on terminal width.

— Object shape named but no concrete worked example given. v0.109.25

└──▷ BREAKING ON UPGRADE

! A supervisor or script that relied on `auth-service --callout-only --spaces` exiting nonzero when one space could not open now sees the host start instead; callers must read the ready line or per-space lines to determine which spaces are retrying.

! `--service-endpoints` now refuses to be used alongside `--spaces`; previously (from v0.112.0) it enabled service endpoints for every lane. Set `"serviceEndpoints": true` on

each space entry in the manifest that requires them and remove the `--service-endpoints` flag; see UPGRADING.

- ! Every multi-space host now writes `<manifest>.applied` before opening any lane and refuses to start if it cannot write that file; the manifest's directory must be writable by the host before upgrading — this affects existing `auth-service --callout-only --spaces` and `deliver --spaces` hosts.
- ! `deliver --spaces` no longer exits nonzero when one space fails to open at startup; supervisors or scripts that relied on that nonzero exit to detect partial failures must instead read the host's ready line or per-space lines for spaces being retried.
- ! `KV_cotal_issued_<space>` and `KV_cotal_accepted_<space>` stores created by any release before v0.111.0 are refused by every binder — stop the space's auth service, callout replicas, and managers; back up both streams with `nats stream backup`; remove them with `nats stream rm`; then start v0.111.0 to recreate them. Generations issued before the upgrade will be refused afterwards.
- ! `--creds`, `--dev-mint`, `--shard`, `--shards`, and `deliver pending` refuse when used with `--spaces`.
- ! `--spaces` requires `--server` to be specified; omitting `--server` with `--spaces` is an error.
- ! `--space` and `--spaces` are mutually exclusive alternatives and cannot be combined.
- ! Foreground `cotal spawn` of Claude Code no longer passes `--permission-mode bypassPermissions`; Claude now keeps its own permission mode, so users who relied on bypass being set automatically on spawn must accept the Bypass Permissions prompt manually or pass `--opt permission-mode=default`.

WAS THIS USEFUL? 👍 👎

Databricks Omnigent

Feedback ⓘ v0.20.0 NOTES ↗ RANK

Open-source meta-harness that orchestrates Claude Code, Codex, Cursor, Pi and custom agents behind one interface, with policy controls, OS-level sandboxing and live session sharing.

> Omnigent v0.20.0 adds full in-tab pull request management, Databricks Sandbox as a managed host type, MCP server visibility via `/mcp`, several composer shortcuts, and two breaking changes to Android OAuth and local server bundle restrictions.

└──▷ **WHAT SHIPPED · 8 FEATURES** most completely described first ? what's the number?

01 **MCP server visibility in composer** NEW 83

Adds `/mcp` composer command to inspect a session's MCP servers and their connection status, with inline sign-in for services and host servers that require authentication.

Check which MCP servers are active and sign in to any that need authentication, without leaving the composer.

📍 `/mcp`

— Named command paired with a runnable usage example. v0.20.0

02 **Android Databricks OAuth callback change** BREAKING 75

Android Databricks workspace sign-in now uses a private OAuth callback: before updating the app, register `ai.omnigent.android://mobile-redirect` for the public OAuth client.

— Exact callback URI and required pre-update action given. v0.20.0

03 **In-tab pull request management** NEW 70

Enables full pull request management from the Pull Requests tab: merge or disable auto-merge, edit title and description, add or remove labels, request or withdraw reviewers, and post comments.

— Names the tab and lists every action; no step-by-step walkthrough. v0.20.0

04 **Composer shortcuts and session links** NEW 70

Adds `/clear` command for web chat in Claude Code and Codex sessions, Shift+Tab to cycle the permission mode in native sessions, and a `?prompt=` link parameter to open a new chat with a prefilled prompt.

Open a new Omnigent web chat session with a question pre-filled, useful for sharing a repeatable prompt link with teammates.

```
$ https://<your-omnigent-host>/?
prompt=Summarize+the+last+10+commits+in+this+repo
```

— Three named shortcuts grouped, one shown via a runnable link example. v0.20.0

05 **Bundle restrictions on local single-user servers** BREAKING 70

Agent uploads and config edits on local single-user HTTP servers now enforce the same bundle restrictions as multi-user servers — bundles with server-side callable or source-file tools, unregistered policy handlers, or absolute or escaping working directories must move to operator startup registration or supported uploaded-agent settings.

— Detailed migration criteria, but no exact command shown. v0.20.0

06 **Databricks Sandbox managed hosts** NEW 65

Supports Databricks Sandboxes as managed hosts, created and run as the user via their Databricks account connection with no SSH and no shared credential.

— Clear mechanism described, but no setup command shown. v0.20.0

THINNER COVERAGE BELOW

07 **File link and shell call rendering in chat** NEW 55

Adds file link rendering that opens line ranges in the format `file.ts:10-20` and displays shell calls in terminal style (`$ command` , output, exit status) in chat.

— Named format but no example of invoking it. v0.20.0

08 **Agent tool to open workspace files** NEW 40

Adds `ui_open_file` agent tool so agents can open workspace files directly in your session.

— Bare description of a named tool, no example. v0.20.0

└──▷ BREAKING ON UPGRADE

! Android Databricks workspace sign-in now uses a private OAuth callback: before updating the app, register `ai.omnigent.android://mobile-redirect` for the public OAuth client.

! Agent uploads and config edits on local single-user HTTP servers now enforce the same bundle restrictions as multi-user servers — bundles with server-side callable or source-file tools, unregistered policy handlers, or absolute or escaping working directories must move to operator startup registration or supported uploaded-agent settings.

WAS THIS USEFUL?  

Graphicle AI Graphicle

Feedback ⓘ v0.1.9 NOTES ↗

RANK

Code graph and memory for your coding agents, in one Rust program. For Claude Code, Codex, Cursor and any MCP client.

> Graphicle v0.1.9 ships four new commands for repo-scale agent workflows — `pack`, `ask --pack`, `plan`, and `fix` — plus Meta Model API support with prompt-cache layouts, a new MCP `pack` tool, and broader multi-language pack ranking.

└──▷ **WHAT SHIPPED · 7 FEATURES** most completely described first ? what's the number?

01 **Automated issue fixing via graphicle fix** NEW 98

`graphicle fix "issue"` (or `--issue-file`) makes a pack and plan, applies search-and-replace edits on a new git branch in a new worktree, runs `--test CMD` (or `fix.test` config) with `--timeout` (1–86,400 seconds) for up to `--rounds` rounds (default 3), and optionally `--commit`s on the new branch. `--focus PATH[:LINE|:START-END]` (up to 50 paths) promotes specific passages to the top of the pack's ranking regardless of the issue's words.

— Full command flow with all flags and defaults v0.1.9

02 **Whole-repo packing via graphicle pack** NEW 95

`graphicle pack [QUESTION]` gathers a project's code for one question into a single text block within a token budget via `--budget` (default 128,000, up to 2,000,000), outputting Markdown, XML, or JSON (`graphicle.pack.v1`). `--layout cache`, available on `graphicle pack`, `graphicle ask --pack`, and `graphicle fix`, splits a pack for a provider's prompt cache, with `--prefix-share F` (default 0.9) controlling the repository part's share of the budget. `--tokenizer tokenizer.json` on `graphicle pack` accepts any Hugging Face tokenizer via the pure-Rust `tokenizers` library for exact token counting instead of byte-per-language estimation.

— Names exact flags, defaults, formats and mechanism v0.1.9

03 **Model profiles and Meta Model API integration** NEW 93

`graphicle config set model meta` writes a Muse Spark profile for the Meta Model API (`https://api.meta.ai/v1`, Responses API, key in `MODEL_API_KEY`, model `muse-spark-1.3`, `reasoning_effort` medium), with `meta:muse-wonder-1.0` selecting Muse Wonder and `compatible:MODEL --url URL` configuring any OpenAI-compatible server. Requests to the Meta Model API now include `prompt_cache_key` and 24-hour `prompt_cache_retention`, with other profiles opting in via `"prompt_cache": true`. A new `POST /responses/input_tokens` integration

counts tokens exactly for requests near the input limit, and model profile fields now cover `input` limits up to 1,048,576 tokens, `output` limits up to 131,072, an optional `price` (USD per million input/output/cached-input tokens), and a `tokenizer` (a `tokenizer.json` path) for exact local counting.

— Names config commands, endpoint and all profile fields v0.1.9

04 **Change plan generation via graphicle plan** NEW 83

`graphicle plan "change request"` requests a plan as JSON (steps, files to edit with line ranges, new files, tests, risks), checks it against the working tree, and prints it as Markdown or JSON without modifying any files.

— Exact command and output schema given v0.1.9

05 **Pack ranking language support expansion** IMPROVED 73

Pack ranking now supports TypeScript, Java, and C++ repositories: it resolves TypeScript/JavaScript imports through relative paths, index files, and `tsconfig` path aliases, and matches file names to code names in Java, TypeScript/JavaScript, and C/C++. Ranking also strips environment sections from issues (headings labelling system info, tested versions, `<name> version` lines, OS/environment fields) so unrelated files are not promoted.

— Describes mechanism but no direct command to invoke v0.1.9

06 **One-shot question answering via graphicle ask --pack** NEW 70

`graphicle ask --pack` answers in one model call from a pack sized to the model's input limit, streaming the answer and verifying every cited `path:line` against the working tree.

— Runnable command with clear verification mechanism v0.1.9

07 **pack tool in MCP server** NEW 67

The MCP server gains a `pack` tool (default 32,000 tokens, at most 200,000), enabled for Claude Code via `graphicle connect claude-code --tools all`.

— Names tool, defaults, and enabling command v0.1.9

WAS THIS USEFUL?  

PydanticAI

Feedback ⓘ v2.63.0 NOTES ↗

PydanticAI is a framework that builds type-safe Python agents with dependency injection, model integrations, tools, and structured outputs.

Adds `IPythonCodeMode`, `DetectLoops` capability, `PostgresSpendStore`, `pydantic_ai.framing` module, `manage_message_history` for UI adapters, and `task_timeout` for `Dataset.evaluate`.

└──▷ GET THIS VERSION ▾

pip

```
$ pip install clai==2.63.0
# already installed? upgrade to this version:
$ pip install --upgrade clai==2.63.0
```

git

```
$ git clone --branch v2.63.0 https://github.com/pydantic/pydantic-ai.git
# already have the repo? check out this version:
$ git checkout v2.63.0
```

PyPI attestation · checked 28 Oct Wrong command?

Adds `task_timeout` parameter to `Dataset.evaluate` to bound each task attempt with a per-task time limit.

Adds `IPythonCodeMode` class alongside the existing Monty-based `CodeMode` for running model-written code in an IPython kernel.

Adds `DetectLoops` capability to interrupt repeated-action loops in agent runs.

Adds `manage_message_history` option to UI adapters so a server-held history is not sent to the model twice.

Adds a public `pydantic_ai.framing` module with `wrap`, `neutralize_tags`, and `register_tag` functions so extensions can frame their own tags.

+5 more

NEW INTEGRATION ag-ui-protocol 1.0 support

NEW SUPPORTED PLATFORM OpenAI gpt-6.1-terra model support

NEW INTEGRATION PostgresSpendStore for shared SpendLimits budgets via PostgreSQL

└──▷ BREAKING ON UPGRADE

- ! `CodeMode` now defaults to `dynamic_catalog=True`; agents relying on the previous behavior of rewriting the cached prompt when a tool is revealed will behave differently.
- ! `GoogleRealtimeConnection.dial` now receives the Gemini Live reconnect replay as a keyword-only `replay` argument; callers passing it positionally will break.
- ! Harness `LLMReminder` now defaults to `interval=10`; setups that relied on reminders being generated on every request must explicitly set `interval` to their previous behavior.
- ! Signed and encrypted history data is now replayed only to providers sharing the same `canonical_system`; cross-provider replay setups will no longer receive history.

WAS THIS USEFUL? 👍 👎

ZeroClaw

Feedback ⓘ v0.8.7 NOTES ↗ RANK

ZeroClaw runs personal AI assistants with configurable model providers, tools, memory, and messaging channels.

> ZeroClaw v0.8.7 patches Wasmtime/TLS security advisories (requiring plugin rebuilds) and swaps out a legacy TLS dependency, while adding cost-warning thresholds, sandbox hardening, native tool-calling support, and several plugin configuration and lifecycle improvements.

└──▷ **WHAT SHIPPED · 8 FEATURES** most completely described first ? what's the number?

01 Sandbox hardening for Firejail and bubblewrap IMPROVED

95

Risk profile `firejail_args` are applied to the Firejail wrapper; entries that would undo ZeroClaw's own hardening, or are not single `--option` / `--option=value` flags, are rejected, causing the sandbox to refuse every command rather than run without them. Separately, when a requested bubblewrap backend is rejected, ZeroClaw now names the cause: a missing or broken launcher, or a build without `sandbox-bubblewrap`.

Restrict a Firejail sandbox to only your approved flags — any flag that weakens ZeroClaw's hardening is rejected and the sandbox refuses all commands rather than running unsafely.

toml

```toml
# In your risk profile block in ~/.zeroclaw/config.toml:
[risk_profiles.<alias>]
firejail_args = ["--net=none", "--noroot"]
```

— Exact config block and rejection behaviour, with runnable example v0.8.7

02 Plugin config and lifecycle CLI improvements IMPROVED

95

`zeroclaw config set` now accepts a non-secret plugin setting directly on the command line (e.g. `zeroclaw config set discord guild_ids 123456789012345678`); only properties the package's `config_schema` marks `x-secret` still use the masked prompt. `plugin remove` now removes a loaded plugin from the directory it was loaded from, instead of deleting `<plugins_dir>/<name>`. `plugin info` and `plugin list --verify` now report a tool whose retained configuration the runtime refuses as `does not load`, along with the runtime's reason and the row holding the values.

Set a non-secret plugin config value (e.g. Discord guild IDs) directly on the command line without a masked prompt.

```
$ zeroclaw config set discord guild_ids 123456789012345678
```

Inspect a plugin that may be failing to load due to retained configuration the runtime rejects.

```
$ zeroclaw plugin list --verify
```

— Runnable commands and exact flags for each fix v0.8.7

03 **Wasmtime and MQTT dependency security patches** BREAKING

75

Patches Wasmtime and TLS advisories; precompiled `.cwasm` plugin components built for Wasmtime versions before 48.0.7 must be rebuilt for Wasmtime 48.0.7. The optional MQTT channel moves from `rumqttc` 0.25 to `rumqttc-v4-next` 0.34 (MQTT 3.1.1), removing the last copy of legacy `rustls-webpki` 0.102.8 and its four audit exceptions.

— Names exact versions and the rebuild requirement, no migration command v0.8.7

04 **native_tools setting for OpenAI-compatible models** NEW

65

Adds `native_tools` support for every OpenAI-compatible chat-completions family: `true` selects native tool calling, `false` selects prompt-guided tools, and unset keeps the family default.

— Names the config key and its three states, no usage example given v0.8.7

05 **Plugin tool name conflict reservation** NEW

60

Reserves `mcp_resources`, `mcp_prompts`, peripheral, and skill-defined tool names before plugin tools register; a conflicting plugin tool is dropped with a `plugin_tool_name_conflict` warning.

— Names the reserved namespaces and warning but no remediation step v0.8.7

06 **Time-bounded Qdrant vector search** IMPROVED

60

Pushes `since` / `until` time bounds into Qdrant vector search so out-of-window high-scoring points no longer consume requested result slots.

— Names the filter fields and effect, no config example v0.8.7

THINNER COVERAGE BELOW

07 **Cost threshold warnings in ZeroCode** NEW 55

Adds `warn_at_percent` cost threshold to ZeroCode, surfaced as a one-time chat line before a session's budget is exhausted.

— Names the config key but gives no example of setting it v0.8.7

08 **Daemon config lifecycle lock relocation** BREAKING 50

The daemon's config lifecycle lock moved to an owner-only `data/.lifecycle` directory; operators should review the Breaking Changes section before upgrading a running daemon.

— Names the new directory but no migration steps given v0.8.7

└──▷ BREAKING ON UPGRADE

! Precompiled `.cwasm` plugin components built for Wasmtime versions before 48.0.7 must be rebuilt for Wasmtime 48.0.7.

! The daemon's config lifecycle lock moved to an owner-only `data/.lifecycle` directory; review the Breaking Changes section before upgrading a running daemon.

WAS THIS USEFUL? 👍 👎

CopilotKit

Feedback ⓘ v1.82.0 NOTES ↗ RANK

CopilotKit connects AI agents to application interfaces with shared state, tools, and generative UI components.

> CopilotKit shipped native generative UI tooling and server-side control over the Automatic Learning pipeline, alongside request tracing and completed i18n coverage for the chat UI.

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

01 **Native generative UI via createGenerativeUI** NEW 75

Adds `createGenerativeUI` with tool and markdown transports, a standard component catalog, and an inspector pane for building native generative UI experiences.

– Names the API and its transports but no code sample v1.81.0

02 **i18n labels for remaining chat strings** IMPROVED 65

Translates the chat's remaining hard-coded strings through a `labels` prop in the `react-core`, `vue`, and `angular` packages, completing i18n coverage for the chat UI.

– Names exact prop and all affected packages v1.82.0

THINNER COVERAGE BELOW

03 **Server-side Learning management methods** NEW 45

Adds server-side Learning management methods to the CopilotKit runtime for programmatic control of the Automatic Learning pipeline.

– Describes purpose but not specific method names v1.81.0

04 **Per-request ID on Intelligence lock calls** IMPROVED 40

Sends one request ID per user request on the Intelligence lock call, enabling better request tracing across the runtime.

– Explains mechanism but no way for reader to act on it v1.82.0

05 **Runnable Vue documentation showcase** NEW 20

Adds a runnable Vue documentation showcase to the example set.

– Bare mention with no details on content or location v1.82.0

WAS THIS USEFUL? 👍 👎

CrewAI

Feedback ⓘ 1.15.31 NOTES ↗ RANK

Open-source Python framework for building multi-agent workflows using role-based AI agent Crews and event-driven Flows.

> CrewAI 1.15.31 adds conversational flow context compaction to help manage long-running multi-agent flow contexts.

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

01 **Conversational flow context compaction** NEW 20

CrewAI now supports conversational flow context compaction to manage long-running flow contexts, aimed at multi-agent workflows that run for extended periods.

— No mechanism, config key, or API detail given 1.15.31

WAS THIS USEFUL? 👍 👎

HyperResearch

Feedback ⓘ v0.13.0 NOTES ↗ RANK

HyperResearch collects and synthesizes web research into a persistent searchable wiki for AI agents.

> HyperResearch v0.13.0 ships full installation and distribution support for OpenAI Codex alongside new run-level spend controls, an async MCP fetch pipeline, and a rebuilt Codex bench harness.

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

01 **Codex installation target for pipeline** NEW 95

`hyperresearch install . --target codex` installs the full 18-step pipeline for OpenAI Codex, writing the entry skill to `.agents/skills/hyperresearch/`, step procedures to `.hyperresearch/codex/steps/`, subagents to `.codex/agents/*.toml`, a Stop hook to `.codex/hooks.json`, and an `AGENTS.md` block. `--target all` installs both Claude Code and Codex side by side, and `--global --target codex` writes to `~/.agents/skills/` and `~/.codex/agents/`.

Install HyperResearch for OpenAI Codex in your project, then launch a sandboxed session with write and network access.

```
$ pip install hyperresearch && hyperresearch install . --target codex
codex --sandbox workspace-write -c sandbox_workspace_write.network_access=true
```

— Exact install command and every target path named v0.13.0

02 **Codex plugin marketplace listing** NEW 90

Adds a Codex plugin marketplace entry via `.codex-plugin/plugin.json` and `.agents/plugins/marketplace.json`, installable with `codex plugin marketplace add jordan-gibbs/hyperresearch` then `codex plugin add hyperresearch@hyperresearch`. The skill is invoked as `$hyperresearch:deep-research` from the plugin, or as `$deep-research` when installed standalone via `npx skills add jordan-gibbs/hyperresearch -a codex`.

— Named config files and exact plugin commands v0.13.0

03 **Run spend limits and reconciliation** NEW 90

`hpr run init --max-sources` and `--max-notes` set ceilings that block pipeline execution at step boundaries. `hpr run step --status done` and `hpr run`

`reconcile` count tagged notes and distinct fetched URLs (HTML and PDFs) in the vault DB.

Cap a research run at 80 sources and 150 notes so the pipeline halts at step boundaries before ballooning costs.

```
$ hpr run init --max-sources 80 --max-notes 150
```

— Exact flags and command with runnable example v0.13.0

04 **Codex bench harness rebuild** IMPROVED 80

Rebuilds the Codex bench harness under `codex_bench/`, running each query in its own git-initialized project via `install --target codex`. `result.json` records vault note count, manifest steps, tier, subagent spawns, sandbox rejections, and `run verify`; a `--dry-run` flag prints the exact `codex exec` command without executing it.

— Named flag and output fields but no full command shown v0.13.0

05 **Async MCP fetch_url** IMPROVED 65

Makes the MCP `fetch_url` tool async so the server handles other requests during a fetch, using a split `prepare_fetch` -> `prov.fetch` (run in a worker thread) -> `save_fetched` path.

— Mechanism named but no reader-facing command v0.13.0

THINNER COVERAGE BELOW

06 **SSRF hardening** IMPROVED 10

The release summary notes SSRF hardening was added, but no further mechanism or scope is detailed in the changelog.

— Mentioned only in summary with no detail v0.13.0

WAS THIS USEFUL? 👍 👎

ag2

Feedback ⓘ v1.2.2 NOTES ↗ RANK

AG2 (formerly AutoGen): The Open-Source AgentOS.Join us at: https://discord.gg/sNGSwQME3x

> ag2 v1.2.2 adds per-call timeout controls for tool calls and callbacks, Anthropic SDK 2.x support with new AnthropicConfig options, and begins deprecating OpenAIConfig ahead of a default switch to the Responses API, alongside an A2UI v1.0 spec update.

└──▷ **WHAT SHIPPED · 7 FEATURES** most completely described first ? what's the number?

01 **Per-call timeouts for tools and callbacks** NEW 88

Adds a `timeout=` parameter to the `@tool` decorator to enforce per-invocation time limits on function tool calls, returning a `TimeoutError` to the model on expiry. The same `timeout=` parameter is added to stream subscribers, `@observer`, `BaseObserver`, `LoopDetector`, and `TokenMonitor` callbacks to limit each callback invocation.

Prevent a slow external-data tool from hanging an agent turn — cap each call at 10 seconds so the model receives a TimeoutError and can continue.

python

```python
from ag2.tools import tool

@tool(timeout=10)
async def fetch_external_data(query: str) -> str:
    # network call that might hang
    ...
```

— Runnable code example, exact decorator and callback names given v1.2.2

02 **OpenAICompletionsConfig and OpenAIResponsesConfig, OpenAIConfig deprecated**

88

BREAKING

Introduces `OpenAICompletionsConfig` as the explicit Chat Completions config class, and `OpenAIResponsesConfig` to opt into the Responses API ahead of the 1.4.0 default change. `OpenAIConfig` is deprecated and will become the Responses API default in 1.4.0; it now emits a `DeprecationWarning` from create(). Switch to `OpenAICompletionsConfig` to keep Chat Completions behavior.

Keep Chat Completions behaviour explicitly after upgrading, before the 1.4.0 default switch to the Responses API.

```python
from ag2.config import OpenAICompletionsConfig

config = OpenAICompletionsConfig(model="gpt-4o-mini")
```

— Exact classes and migration path given with runnable example v1.2.2

03 **AnthropicConfig options and Anthropic SDK 2.x support** NEW

81

Adds an `open_telemetry` option to `AnthropicConfig` to configure or disable the Anthropic SDK's own OpenTelemetry spans, and a typed `speed` option to control Anthropic inference speed. Also adds support for Anthropic SDK 2.x (requires `anthropic>=1.15.0,<3`).

Disable the Anthropic SDK's built-in OpenTelemetry spans to avoid double-tracing when your own instrumentation is already in place.

```python
from ag2.config import AnthropicConfig

config = AnthropicConfig(
    model="claude-opus-4-5",
    open_telemetry=False,
)
```

— Config fields named with a runnable example, version requirement given v1.2.2

04 **Deprecated sampling params on AnthropicConfig** DEPRECATED

68

`temperature`, `top_p`, and `top_k` on `AnthropicConfig` are deprecated: with `anthropic` SDK 2.x they are silently dropped since the API rejects them; with 1.x they still warn but are still sent.

— Names exact fields and before/after behaviour, no migration example v1.2.2

THINNER COVERAGE BELOW

05 **question_instructions field in OpenAIDecisionsConfig** NEW

50

Adds a `question_instructions` field to `OpenAIDecisionsConfig`, accepting strings, ag2 inputs, or SDK parts, which is appended to the agent prompt.

— Field named but no usage example provided v1.2.2

06 **A2UI v1.0 spec update** IMPROVED 50

Updates A2UI v1.0 to the current spec, including `callAgentFunction` / `agentFunctionResponse` and renderer capabilities.

— Names updated spec methods but no usage detail v1.2.2

07 **Updated OpenAI SDK version requirement** BREAKING 35

Now requires `openai>=3.37.0,<4`.

— Bare version constraint with no further context v1.2.2

└──▷ **BREAKING ON UPGRADE**

! `OpenAIConfig` is deprecated and will become the Responses API default in 1.4.0; it now emits a `DeprecationWarning` from create(). Switch to `OpenAICompletionsConfig` to keep Chat Completions behavior.

! `temperature`, `top_p`, and `top_k` on `AnthropicConfig` are deprecated: with `anthropic` SDK 2.x they are silently dropped (the API rejects them); with 1.x they warn but are still sent.

! Now requires `openai>=3.37.0,<4`.

! Now requires `anthropic>=1.15.0,<3`.

WAS THIS USEFUL? 👍 👎

Agno (formerly Phidata)

Feedback ⓘ v3.1.11 NOTES ↗

RANK

Agno is an agent framework and runtime that orchestrates and runs multi-agent systems.

> Agno v3.1.11 adds a session search API with rich filtering, expands Company Task and Routine lifecycle management with new endpoints, and lets ClaudeAgent and CodexAgent run as full members of Teams and Workflows.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 **Session search and filter-schema API** NEW 95

Adds `GET /sessions/filter-schema` and `POST /sessions/search` endpoints supporting nested AND/OR/NOT filters on session name, session/user/component IDs, session type, and created/updated timestamps. PostgreSQL and SQLite get predicate pushdown for these filters, with a batched fallback for other adapters, and both sync and async client methods are included.

— Full endpoint names, filter mechanism, and backend behaviour given v3.1.11

02 **ClaudeAgent and CodexAgent as Team/Workflow members** NEW 90

`ClaudeAgent` and `CodexAgent` can now serve as members of an Agno Team and as steps of a Workflow, including inside Parallel, Loop, Condition, and Router constructs. Each member gets per-member model/tools/workspace isolation and per-session conversation persistence, with optional `sandbox=` container execution, and AgentOS lists harness members on the team page.

— Names constructs and sandbox flag; UI location given but no exact steps v3.1.11

03 **Company Task and Routine lifecycle management** NEW 85

`POST /company/tasks` now accepts a `status` field so tasks can be created directly in Backlog, Ready, or Cancelled states, with free movement between stages and the ability to reopen Done and Cancelled tasks. New routine endpoints `POST /company/routines/{id}/cancel`, `POST /company/routines/{id}/archive`, `POST /company/routines/{id}/restore`, and an admin-only `DELETE /company/routines/{id}` were added, and `GET /company/routines` gains `sort_by` and `sort_order` query parameters.

— Every endpoint and parameter named verbatim, grouped by module v3.1.11

WAS THIS USEFUL? 👍 👎

atomic-agent

Feedback ⓘ v0.6.12 NOTES ↗ RANK

Atomic Agent is a local-first AI agent. Runs open-weight models on your own machine via llama.cpp.

> atomic-agent shipped key-protected remote access with a `/panel` for managing chats from any device, native Linux packaging (deb, rpm, AUR) with package-manager-aware updates, and live model switching without restarts, alongside a new `atag doctor` diagnostic command, chat folders, and per-chat working directories.

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

01 **atag doctor diagnostic command** NEW 85

`atag doctor` (alias `atomic-agent doctor`) checks the engine, port, model, memory, keys, and connectivity, and fixes what it can automatically.

Diagnose a broken local model or connectivity issue in one command before filing a bug report.

```
$ atag doctor
```

— Exact command named, example given, mechanism listed v0.6.12

02 **Linux native packaging and package-manager-aware updates** NEW 80

Ships `atomic-agent-bin` and `atomic-agent-desktop-bin` on the AUR, plus `deb` and `rpm` packages from a signed repository, and `apt` / `dnf` packages, enabling system-package-manager installs and updates on Linux. `atomic-agent update` and `atomic-agent uninstall` now detect package-manager installs and redirect users to their package manager instead of self-updating.

— Names every package type and command, no install walkthrough desktop-v0.0.6 v0.6.12

03 **Agent menu enhancements** NEW 78

Adds an Agent menu option to connect to another computer by address and key and use its models (with the agent server now requiring a key when exposed on a network address, and its `/panel` page providing chats, replies, and approvals from any device), plus an Agent menu option to keep the agent running after the window closes.

— Names key requirement, /panel, and Agent menu option Names exact menu option, minimal further detail

desktop-v0.0.6 v0.6.12

04 **Live model switching without restart** IMPROVED 60

Model and config can be switched live without restarting the agent: first via the local API, then in the desktop app from the composer chip or `Settings › Models`.

– Clear UI path but no command or endpoint named `desktop-v0.0.6` `v0.6.12`

THINNER COVERAGE BELOW

05 **Agent data backup, restore, move and reset** NEW 50

Adds backup, restore, move, and reset of agent data, with reset preserving downloaded models.

– Describes behavior but no command or path given `v0.6.12`

06 **Chat folders in sidebar** NEW 50

Adds chat folders to the sidebar with drag-and-drop organization; imported chats preserve the folders they came with.

– UI feature described but no exact navigation path `desktop-v0.0.6`

07 **Per-chat working folder** NEW 50

Adds a per-chat working folder that can be changed without restarting the agent or stopping a background task.

– Explains behavior but no setting location named `desktop-v0.0.6`

08 **Exa key in integrations** NEW 40

Adds an Exa key field in `/integrations`.

– Named surface but no further mechanism described `v0.6.12`

09 **Background starter-model download** NEW 40

Adds background starter-model download on first run so users can start typing immediately with no model present yet.

– Clear behavior, no model name or size given `desktop-v0.0.6`

10 **Read-aloud button on replies** NEW 30

Adds a read-aloud button on chat replies on macOS.

– Bare feature name, no further mechanism described `desktop-v0.0.6`

WAS THIS USEFUL? 👍 👎

AutoGPT

Feedback ⓘ autogpt-platform-beta-v0.10.0 NOTES ↗ RANK

AutoGPT automates multi-step workflows with continuous AI agents that users can build and deploy.

> AutoGPT Platform beta v0.10.0 introduces External API v2 (deprecating v1), hardens self-hosted Docker Compose installs with per-install service passwords and localhost-only service binding, and adds a shadow-mode read-injection check alongside new models, limits and integration tweaks.

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

01 Self-hosted Docker Compose security hardening BREAKING

80

Adds `make init-env` command for self-hosted Docker Compose installs to generate per-install service passwords; PostgreSQL, RabbitMQ, Redis, FalkorDB, and ClamAV now bind to `127.0.0.1` only instead of all interfaces. Existing installs must run `make init-env` (or the installer) before `docker compose up`, following the 'Upgrading: service passwords are generated per install' migration guide — it will not start otherwise.

Bootstrap a fresh self-hosted install so Docker Compose can generate and load per-install service passwords before first launch.

```
$ make init-env && docker compose up
```

— Names command, affected services, and gives migration path

autogpt-platform-beta-v0.10.0

THINNER COVERAGE BELOW

02 Shadow mode for read-injection check NEW

50

Adds a dedicated flag and shadow mode to the read-injection check — shadow mode judges every read and holds nothing, enabling zero-impact auditing.

— Explains mechanism but flag name not given autogpt-platform-beta-v0.10.0

03 New configurable safety limits NEW

50

Adds new configurable limits: caps HTTP response body reads under a configurable size limit, caps field sizes a template can format, and runs every user-supplied regex pattern under a time limit.

— Names three limits but no config keys or defaults autogpt-platform-beta-v0.10.0

04 **Airtable base picker and change preview** NEW — 50

Pick Airtable bases from a card and preview table changes before committing.

— Describes UI flow with a clear starting point `autogpt-platform-beta-v0.10.0`

05 **External API v2 released, v1 deprecated** BREAKING — 45

External API v2 supersedes v1; v1 remains active until 2027-01-31 before being retired.

— Names versions and deprecation date but no endpoint details

`autogpt-platform-beta-v0.10.0`

06 **New model support: Kimi K3.1 Pro, GPT-6.1 Terra** NEW — 30

Adds `Kimi K3.1 Pro` and `GPT-6.1 Terra` model support.

— Names models but no usage specifics `autogpt-platform-beta-v0.10.0`

07 **Cloudflare added to MCP catalog** NEW — 25

Adds Cloudflare to the MCP catalog.

— Bare addition with no further detail `autogpt-platform-beta-v0.10.0`

08 **API key last-used tracking** NEW — 25

Records last-used timestamp for each API key.

— Describes feature but no UI/API access path given `autogpt-platform-beta-v0.10.0`

09 **Secret-content warning in chat** NEW — 25

Warns users before sending content that looks like a secret in chat.

— Behaviour described without detection mechanism `autogpt-platform-beta-v0.10.0`

10 **Otto model calls tagged with turn trigger** NEW — 25

Tags every Otto model call with the originating turn trigger.

— Minimal description of a tracing feature `autogpt-platform-beta-v0.10.0`

└──▷ BREAKING ON UPGRADE

! Self-hosted Docker Compose now requires running `make init-env` (or the installer) before `docker compose up` — it will not start without generated per-install service passwords. Existing installs must follow the 'Upgrading: service passwords are generated per install' migration guide.

! PostgreSQL, RabbitMQ, Redis, FalkorDB, and ClamAV now listen on `127.0.0.1` only instead of all interfaces.

! External API v1 is deprecated in favour of v2; v1 keeps working until 2027-01-31.

WAS THIS USEFUL? 👍 👎

Cloudflare Agents

Feedback ⓘ agents@0.31.0 NOTES ↗ RANK

The Cloudflare Agents SDK builds stateful agents on Workers and Durable Objects, with persistent state, scheduling, WebSockets, and MCP server support.

> Cloudflare Agents 0.31.0 reworks browser automation with internal-hostname routing and direct session control, moves browserTool onto its own Dynamic Worker (a breaking change), makes MCP SDKs optional peer dependencies, and promotes the AI SDK and Pi AI model adapters to stable.

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

01 browserTool runs in its own Dynamic Worker

BREAKING

90

`browserTool` now runs model code in its own Dynamic Worker instead of on codemode, removing the need for `ctx` or a `CodemodeRuntime` export. The `calls` log output and the `restarted` / `newTabs` result fields are replaced with a `notice` field, and model code inside `browserTool` now uses a `browser` global (browser.cdp(...), browser.spec({ domain })) instead of cdp.send(...). Migration requires removing `ctx` from `browserTool` options, removing the `CodemodeRuntime` export, updating code that read `restarted` / `newTabs` / `calls` to read `notice` instead, and updating any system prompts referencing cdp.send(...). `PiBrowserToolDetails` is reduced to only `executionId` and `status` fields.

— Full before/after mechanism with concrete migration steps agents@0.31.0

02 Internal hostname routing and session control for browser

NEW

88

Adds `internalHosts` parameter to browserRun() (up to 8 hostnames) so the browser routes requests to named internal hosts — such as `preview.internal` — through the agent's `fetch` handler or another Worker instead of the public internet. Adds browser.assignHost() to let the model dynamically point a host with a `target` schema at a container port or workspace folder at runtime, and adds this.browser.session() on the agent so user code can drive the same browser that `browserTool` is using.

Let the model point an internal hostname at a specific container port at runtime, then drive that same browser from your own agent code via session().

```typescript
// Inside your Agent class
async onChat(message) {
  // Model can call browser.assignHost() to pick the target
  // Your code can also drive the browser directly:
  const session = await this.browser.session();
  await session.cdp({ method: 'Page.navigate', params: { url:
'http://workspace.internal' }, sessionId: 'active' });
}
```

— Names exact methods, a hostname limit, and includes a runnable example `agents@0.31.0`

03 **MCP peer dependencies made optional and client import path changed**

78

BREAKING

`@modelcontextprotocol/sdk` and `@modelcontextprotocol/server` are now optional peer dependencies, so projects that don't serve MCP no longer install them. Projects that use `agents/mcp`, `agents/x402`, or `agents/experimental/webmcp` must now explicitly add `@modelcontextprotocol/sdk@1.32.1` and/or `@modelcontextprotocol/server@2.3.1` to their `package.json` or builds fail with an unresolved import. MCP client helpers such as `RPCClientTransport` must now be imported from `agents/mcp/client` instead of `agents/mcp`.

— Names exact packages, versions, and new import path `agents@0.31.0`

04 **Verified auth token passthrough in createMcpHandler** NEW

65

`createMcpHandler` now reads the verified token from `@cloudflare/workers-oauth-provider` 1.x (`ctx.auth`) and passes it to tool callbacks as `context.http.authInfo`, including token, client ID, scopes, expiry, and resource.

— Names the exact field and source but no usage example `agents@0.31.0`

THINNER COVERAGE BELOW

05 **AI SDK and Pi AI model adapters promoted to stable** IMPROVED

35

`agents/models/ai-sdk` and `agents/models/pi-ai` are promoted from experimental to stable.

— States the change but no behavioural detail `agents@0.31.0`

06 **PiWhenBusy reject option** IMPROVED

25

`PiWhenBusy` now accepts the `'reject'` value.

— Bare mention of a new accepted value, no context `agents@0.31.0`

└──▷ BREAKING ON UPGRADE

! Projects that use `agents/mcp`, `agents/x402`, or `agents/experimental/webmcp` must now explicitly add `@modelcontextprotocol/sdk@1.32.1` and/or `@modelcontextprotocol/server@2.3.1` to their `package.json` or builds fail with an unresolved import.

! MCP client helpers such as `RPCClientTransport` must now be imported from `agents/mcp/client` instead of `agents/mcp`.

! `browserTool` users must remove `ctx` from `browserTool`'s options and remove the `CodemodeRuntime` export from their Worker (unless `createBrowserTools` still uses it).

! The `browserTool` result fields `restarted`, `newTabs`, and `calls` are removed; code that inspects results must read `notice` instead.

! `PiBrowserToolDetails` now has only `executionId` and `status`; the previous fields are gone.

! Model code that references cdp.send(...) (e.g. in system prompts) must be updated to browser.cdp(...).

WAS THIS USEFUL?  

OpenClaw

Feedback ⓘ v2026.10.8 NOTES ↗ CODE ↗ RANK

OpenClaw is a personal AI assistant platform that connects messaging channels, tools, and model providers through a gateway.

> OpenClaw added Gemini 4 Argon as a supported model and shipped a cluster of security hardening fixes covering exec allowlisting, sandboxing, proxy trust, and credential leak prevention.

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

01 **Exec allowlist, sandbox, and proxy trust hardening** IMPROVED

54

Exec allowlist mode now requires approval for `test -v` operand expansion, closing a bypass path. Generic node invoke refuses browser proxy commands, and sandboxed spawned children cannot run host CLI backends. The secret egress proxy rejects distrusted eMudhra certificate chains. Fish Audio and Gradium errors no longer expose API keys in logs or error output.

— Names specific bypass paths and providers but no flags or config to act on. v2026.10.8

02 **Gemini 4 Argon provider support** NEW

30

Adds Gemini 4 Argon as a supported Google provider model.

— Names the model but no setup or config detail given. v2026.10.8

WAS THIS USEFUL? 👍 👎

A2A Python SDK

Feedback ⓘ v1.3.0 NOTES ↗ RANK

The A2A Python SDK implements the Agent2Agent protocol for discovering agents and exchanging tasks and messages.

> A2A Python SDK v1.3.0 adds a configurable timeout to automatically release parked interrupted tasks on the server.

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

01 **Configurable timeout for parked task release** NEW 40

The server now supports a configurable timeout that automatically releases parked (interrupted) tasks, preventing them from stalling indefinitely.

— No config key or flag name given, just a described mechanism. v1.3.0

WAS THIS USEFUL? 👍 👎

magic-context

Feedback ⓘ dashboard-v0.20.21 NOTES ↗ RANK

Unbounded context. Memory that manages itself.

> Magic-context's dashboard now parses a previously unrecognized log record type so the full agent log remains visible.

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

01 **Dashboard support for output-reservation log records** IMPROVED 50

Dashboard 0.20.21 adds support for the `output-reservation/v1` Broca log record type, fixing an issue where the dashboard would halt log parsing at that record and hide subsequent steps in the log viewer.

— Names exact record type and bug fixed, but no UI path given dashboard-v0.20.21

WAS THIS USEFUL? 👍 👎

◆ MCP Servers & Tooling

api-mcp-server

Feedback ⓘ v3.18.0 NOTES ↗ RANK

MCP server that exposes Hostinger's platform API for hosting deployment, CDN control, domain management, and WordPress operations.

> Hostinger's MCP server gained hosting website settings and cache controls, new Reach integration operations for ecommerce and AI Website Builder connections, an email template retrieval operation, and bundled agent skill files for guided deployment workflows.

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

01 Website settings retrieval, update, and cache clearing NEW

90

Adds `hosting_websites_get-settings` to retrieve website SSL, CDN, and cache settings, `hosting_websites_update-settings` to update website HTTPS redirect, CDN, and cache settings, and `hosting_cache_clear` to clear website cache on demand.

— Names three callable operations and the exact settings each touches. v3.18.0

02 Reach integrations for ecommerce and AI Website Builder NEW

85

Adds `reach_integrations_connect-ecommerce-store` to connect an ecommerce store via the Reach integrations surface, and `reach_integrations_disconnect-ai-website-builder-site` to disconnect an AI Website Builder site. Related operations `reach_integrations_list-ecommerce-stores` (to audit active integrations) and `reach_integrations_connect-ai-website-builder-site` are also usable in this surface.

List all ecommerce stores connected to a Reach account to audit active integrations.

```
$ reach_integrations_list-ecommerce-stores
```

Connect an AI Website Builder site to a Reach campaign workflow.

```
$ reach_integrations_connect-ai-website-builder-site
```

— Four named operations with runnable commands shown in examples. v3.12.0 v3.18.0

03 **Bundled agent skills for deployment workflows** NEW

60

Bundles agent skills `skills/deploy-to-hosting/SKILL.md` and `skills/headless/steps/08-deploy/SKILL.md` for structured deployment workflows, and adds 4 server instruction files to guide agent behavior.

— Names exact skill file paths but no usage command shown. v3.13.0

THINNER COVERAGE BELOW

04 **Email template retrieval operation** NEW

45

Adds `reach_templates_email` operation to retrieve an email template.

— Single named operation, no further mechanism described. v3.12.0

WAS THIS USEFUL? 👍 👎

whodb

Feedback ⓘ 0.138.0 NOTES ↗ RANK

Where data access meets operational intelligence

> whodb 0.138.0 focused on polish and reliability, updating CLI support, adding a better 404 page, hardening multipart upload and connection probe handling, and refreshing translations.

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

01 **Hardened multipart upload and connection probe handling**

IMPROVED 30

The release hardens multipart upload and connection probe handling, improving robustness of these internal mechanisms, though no specifics on the vulnerability or fix are given.

— Names the two areas but no mechanism detail 0.138.0

02 **CLI support updates** IMPROVED 20

The release updates CLI support with CLI feature fixes, though the specific commands or flags changed are not detailed.

— No specific commands or flags named 0.138.0

03 **Improved 404 page** IMPROVED 20

whodb 0.138.0 adds a better 404 page, per the release summary.

— Mentioned only in passing, no detail on change 0.138.0

04 **Updated translations** IMPROVED 5

Translations were updated in this release.

— Bare mention with no specifics 0.138.0

WAS THIS USEFUL? 👍 👎

mcp-server-plugin

Feedback ⓘ 0.222.v9d40a_8d624e4 NOTES ↗

RANK

> mcp-server-plugin changed the default output of its `getTestResults` tool to reduce noise for MCP clients.

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

01 **Summarized output from getTestResults tool** BREAKING

55

The `getTestResults` tool now returns a short summary by default instead of full test result data, reducing noise for MCP clients.

— Names the exact tool and behavior change but no migration detail for full data

0.222.v9d40a_8d624e4

WAS THIS USEFUL? 👍 👎

fablecut

Feedback ⓘ v1.18.0 NOTES ↗ RANK

Zero-dependency browser video editor that AI agents can drive — JSON timeline, MCP + REST, live-reloading UI

> FableCut v1.18.0 ships three new agent-driveable editing capabilities — beat-synced cutting, subject-following reframe for multi-ratio copies, and automatic highlight reels — each exposed both as a callable tool and through dedicated Inspector/menu UI.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 **Beat detection and beat-synced cutting** NEW 88

The new `fablecut_beats` agent tool detects tempo, fits a beat grid, finds the drop, marks every beat, snaps cuts to the nearest beat on targeted tracks, or re-cuts clips to 1, 2, 4, 8, or 16 beats; each operation is one undo step and preserves a version before snapping or re-cutting. The same capability is surfaced in the UI via a new Inspector Beats section on audio/video clips with Mark, Snap cuts, and Re-cut controls.

— Names tool, UI controls and beat counts; no usage example given v1.18.0

02 **Reframe for 9:16, 1:1, 4:5 copies** NEW 88

The new `fablecut_reframe` agent tool produces a 9:16, 1:1, or 4:5 copy of an edit using a virtual camera that follows the subject in each shot (reacting to motion, detail and colour, held in a dead zone, eased, no faster than a pan speed), writing results as ordinary position keys while leaving hand-placed titles and clips untouched. The UI adds a Reframe section in the Inspector and a Follow the subject in each shot option in New sequence… .

— Detailed mechanism and exact UI/tool names; no runnable example shown v1.18.0

03 **Automatic highlight reel generation** NEW 82

The new `fablecut_highlights` agent tool scores every half-second of the timeline for movement, audio prominence, and picture clarity, then assembles the best moments (15–90 s, any canvas, optionally following the subject) back-to-back into a new sequence in story order. The UI exposes this via Highlights… in the sequence menu.

— Scoring mechanism and duration range named; no sample output given v1.18.0

WAS THIS USEFUL? 👍 👎

gridctl

Feedback ⓘ v1.0.0-rc.3 NOTES ↗ RANK

Gridctl is a command-line tool that manages and routes Model Context Protocol servers, skills, and tools.

> gridctl's rc.3 cycle adds deferred tool discovery to exceed MCP's 100-tool-per-server cap, external secret providers, OAuth resource-server mode and per-tool execution policy blocks, alongside new skill-registry commands and Skills-over-MCP exposure for upstream clients.

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

01 **Skill registry, authoring, and MCP exposure** NEW 83

Adds `gridctl skill add` to register a skill from a git repo or local directory into the registry, `gridctl skill import --kind all` (with `--kind`, local-directory import, and client-home import) to import skills, and `gridctl activate` to promote a draft skill to active/policy-visible status, plus support for authoring skills as `SKILL.md` files in any editor or the Library workspace; active, policy-visible registry skills are now exposed to upstream MCP clients as prompts, resources, and through Skills over MCP (`io.modelcontextprotocol/skills`), including file manifests and directory browsing.

Register a local skill directory and activate it for policy-visible MCP exposure in one workflow.

```
$ gridctl skill add ./my-skill
gridctl activate
```

Import all skills at once — including from local directories and the client home — when bootstrapping a new environment.

```
$ gridctl skill import --kind all
```

— Names every subcommand and flag with runnable examples Explains exposure mechanism but gives no command to try it

product docs

02 **Upstream gateway tool discovery and auth** NEW 83

Adds the `gateway.code_mode: on` config key to `~/.gemini/config/mcp_config.json` so a large MCP tool stack can exceed the 100-tool-per-server cap, `search_tools` and `call_tool` for deferred tool discovery without requiring code execution, a `tool_discovery.mode: deferred` setting for

when the MCP host should not execute code, and OAuth resource-server mode for the upstream gateway.

Enable deferred tool discovery when the MCP host should not execute code, avoiding the 100-tool cap with a large server stack.

```yaml
gateway:
  tool_discovery:
    mode: deferred
```

— Names config keys and functions with a working config example Bare one-line mention with no mechanism or config shown

v1.0.0-rc.3

03 **External secret providers for stack secrets** NEW 75

Adds external secret providers supporting `exec` and 1Password references for runtime secret resolution, letting a stack resolve values like `op://vault/item/field` via a configured `providers` entry instead of embedding them.

Resolve a secret at runtime from a 1Password reference rather than embedding a value in the stack file.

```yaml
secrets:
  sets:
    - dev
  providers:
    - type: exec
      command: [op, read, 'op://vault/item/field']

mcp-servers:
  - name: github
    transport: stdio
    env:
      GITHUB_PERSONAL_ACCESS_TOKEN:
${var:GITHUB_PERSONAL_ACCESS_TOKEN}
```

— Names exec/1Password mechanism with a runnable stack.yaml example v1.0.0-rc.3

04 **Per-tool execution policy block** NEW 75

Adds a per-tool execution `policy:` block with a `block` action to gate individual tool calls, settable per tool entry in a stack file.

Block a specific tool from being called by any client by adding a `policy:` block to its entry in the stack file.

```yaml
mcp-servers:
  - name: github
    image: ghcr.io/github/github-mcp-server:v1.12.1
    transport: stdio
    tools:
      - name: delete_repository
        policy:
          block: true
```

— Named config key with runnable stack.yaml example `v1.0.0-rc.3`

05 **Stack reproducibility and compatibility** NEW 65

Adds pack composition with a committed pack lock for repeatable pack installs, and a `requires.gridctl` floor in stack files to enforce a minimum compatible gridctl version, failing fast on an incompatible host.

Pin the minimum gridctl version a stack requires so it fails fast on an incompatible host.

```yaml
version: '1'
name: daily
requires:
  gridctl: v1.0.0-rc.3
```

— Names the lock concept but no file or command detail Named config key with a runnable stack.yaml example

`v1.0.0-rc.3`

06 **OpenTelemetry tracing and OTLP metrics export** NEW 60

Adds an `otel-collector.yaml` example for exporting traces and metrics to a local OpenTelemetry collector, plus OpenTelemetry semantic conventions and OTLP metrics export.

— Names the example file and OTLP but no exact command `v1.0.0-rc.3`

THINNER COVERAGE BELOW

07 **Headers support and stricter stack validation** IMPROVED 55

Adds `headers:` support to stack validation, along with strict unknown-key enforcement, hidden-tool references, and custom headers.

— Names the validation surfaces but no example config `v1.0.0-rc.3`

08 **Streamable HTTP endpoint for Antigravity 2.0** BREAKING

50

Supports the `/mcp` endpoint (streamable HTTP) for Antigravity 2.0, which is distinct from and replaces the legacy `/sse` endpoint as the correct target.

— Names both endpoints but gives no migration command `product docs`

09 **Retained binary and state-aware rollback** NEW

45

Adds retained binary rollback and state-aware rollback, completing a three-part rollback series.

— References a rollback series without commands or scope `v1.0.0-rc.3`

10 **Bidirectional subscriptions/listen support** NEW

30

Adds `subscriptions/listen` support on the stateless generation in both directions.

— Named protocol method but no usage shown `v1.0.0-rc.3`

└──▷ **BREAKING ON UPGRADE**

! Antigravity 2.0 uses the `/mcp` endpoint for streamable HTTP — the `/sse` endpoint is no longer the correct target.

WAS THIS USEFUL?  

mcp-server

Feedback ⓘ v2.4.0 NOTES ↗ RANK

BrowserStack's Official MCP Server

> BrowserStack MCP Server v2.4.0 adds experimental accessibility capabilities and expands infra product support through the capability registry.

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

01 **Experimental accessibility rules and automated setup** NEW

35

Adds experimental accessibility rules and automated accessibility setup, along with a failed-pages comparison feature in scan results.

— Names three additions but no mechanism, flags, or usage path given. v2.4.0

02 **Infra products onboarded into capability registry** NEW

17

Onboards infra products into the capability registry along with the mechanisms they required.

— Vague prose; no named products, endpoints, or config surfaces. v2.4.0

WAS THIS USEFUL? 👍 👎

kordoc

Feedback ⓘ v4.24.13 NOTES ↗ RANK

모두 파싱해버리겠다 — HWP·HWPX·PDF·Office 문서를 Markdown으로. 양식 자동 채우기와 신구대조를 갖춘 CLI·MCP 서버 | Convert Korean documents (HWP, HWPX, PDF, Office) to Markdown — CLI and MCP server with form filling and diff

> kordoc ships agency-pack (`.kpack`) support, letting institutions bundle and reuse HWPX house-style templates across the CLI, MCP server, and API.

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

01 Agency-pack (`.kpack`) templates across CLI, MCP, API NEW

95

Adds `kordoc pack build` and `kordoc pack info` CLI subcommands to create and inspect `.kpack` agency-pack bundles from a set of institutional HWPX documents, plus a `generate --pack` flag to generate new documents conforming to a `.kpack` template. The same capability is exposed to AI agents via MCP tools `build_agency_pack`, `agency_pack_info`, and `generate_document pack_path`, and programmatically via API methods `buildAgencyPack`, `loadAgencyPack`, and `packTemplate`. The `.kpack` bundle itself is a code-free zip containing `pack.json`, `guide.md`, and `kinds/*.hwpx`, and is treated as untrusted input on load.

Build an agency pack from a folder of institutional HWPX documents, then inspect the result before using it.

```
$ npx kordoc pack build ./서식/*.hwpx -o 우리기관.kpack
npx kordoc pack info 우리기관.kpack
```

Generate a new document that conforms to your agency's house style defined in a `.kpack` bundle.

```
$ npx kordoc generate 보고서.md --pack 우리기관.kpack -o 보고서.hwpx
```

— Names every subcommand, flag, MCP tool, API method and file format with runnable example

v4.24.13

WAS THIS USEFUL? 👍 👎

funplay-cocos-mcp

Feedback ⓘ v0.6.9 NOTES ↗ RANK

The Most Advanced MCP Server for Cocos Creator Editor with `execute_javascript`, prompts/resources, screenshots, asset dependency validation, tool profiles, one-click client configuration, and editor automation.

> funplay-cocos-mcp v0.6.9 adds integration with DeepSeek Harness (DSH), introducing a project-local Cordis overlay and extending Skills tooling to manage DSH's nearest-Git-root skills directory.

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

01 DeepSeek Harness (DSH) integration and skills management NEW

82

Adds DeepSeek Harness (DSH) support via a dedicated project-local Cordis overlay using the official Streamable HTTP MCP client, surfacing an explicit `dsh web --patch` launch command in the dashboard; the overlay is not auto-loaded and existing global/profile YAML is never edited. Configure + Skills, Project Skills, and MCP instruction/Skill tools are extended to support DSH's nearest-Git-root `.dsh/skills` directory, including managed references, update notices, backups, and restore, while preserving existing or locally edited Skills.

Launch the project-local DSH Cordis overlay after configuring it through the Funplay dashboard — use the command the dashboard surfaces to start the Streamable HTTP MCP client without touching global YAML.

```
$ dsh web --patch
```

— Names exact command and directory but lacks config file details v0.6.9

WAS THIS USEFUL? 👍 👎

hister

Feedback ⓘ v0.21.0 NOTES ↗ RANK

Your own search engine

> hister v0.21.0 adds live config reload, three new import sources (Kiwix, Sitemap, Readwise), remote and offline user-management commands, and an expanded search query syntax with boolean operators and new filters, alongside breaking changes to authentication, upload limits, symlink indexing, and MCP client behavior.

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

01 Kiwix, Sitemap, and Readwise import commands NEW

90

`hister import zim collection.zim` (with `--base-url` for URL recovery) indexes HTML and plain text from Kiwix archives; `hister import sitemap FILE_OR_URL` (supporting sitemap indexes, gzip files, stdin, and `--force` to re-index existing documents) indexes pages from XML sitemaps; `hister import readwise` imports Readwise Reader documents including saved content, notes, summaries, tags, and source metadata.

Index a Kiwix archive (e.g. a Wikipedia ZIM file) so its articles become searchable without re-downloading each page.

```
$ hister import zim collection.zim --base-url
https://en.wikipedia.org
```

— Three import commands with exact flags listed. v0.21.0

02 Remote and offline user management commands BREAKING

90

`hister user create`, `show`, `update`, and `delete` manage accounts on a running server using an administrator's personal token; `hister user create USERNAME --offline --admin` creates the first administrator or recovers access with the server stopped. The old `create-user`, `show-user`, `update-user`, and `delete-user` names are now deprecated aliases, the old token-display flag is replaced by `--show-token`, and `--token` now supplies authentication credentials.

— Exact commands, flags, and deprecated aliases named. v0.21.0

03 Expressive search query syntax NEW

90

`site:` filters now include subdomains, and a new `has:` filter finds documents with a populated field or metadata value (e.g. `has:label`, `has:metadata.author`). Search

now supports nested group expressions and `AND`, `OR`, `NOT` boolean operators, with invalid queries explaining what needs fixing and where. Uppercase `AND`, `OR`, and `NOT` are now reserved operators (quote them for literal searches), negation takes precedence over `AND`, then `OR`, and malformed queries are rejected by search, update, and delete operations.

— Filters, operators, precedence and rejection behavior all named. v0.21.0

04 **Language detection and pinning controls** NEW 85

`indexer.languages` limits automatic language detection to a set of languages, `indexer.language_detection_accuracy` chooses between `low` (memory) and `high` (accuracy) modes, and `hister update QUERY --language en` pins a document's language across reindexing, imports, and updates (`--language auto` clears a pin, `--language unknown` keeps a document without an identified language). Run `hister reindex` to apply detection changes to existing documents.

Enable proxy-based SSO and tune language detection accuracy in the server config.

yaml

```yaml
server:
  proxy_auth_header: X-Remote-User

indexer:
  languages:
    - en
    - de
  language_detection_accuracy: high
```

— All config keys and flag values named. v0.21.0

05 **Live config reload without restart** NEW 80

`hister reload` applies updated configuration, rules, and logging changes to a running server without a manual restart, and rejects invalid settings before applying them.

Apply updated config, rules, or logging changes on a live server without a full restart.

```
$ hister reload
```

— Exact command and validation behavior given. v0.21.0

06 **Proxy and OIDC identity configuration** BREAKING 80

`server.proxy_auth_header` enables sign-in through a trusted authentication proxy, and `identity_claim` lets OIDC providers select a stable account identity for

providers without email addresses. Changing `identity_claim` requires relinking existing accounts (leaving it unset preserves the existing email-based identity); existing GitHub account links are not automatically migrated to permanent account IDs, so admins should back up the database and arrange verified relinking before users sign in.

Enable proxy-based SSO and tune language detection accuracy in the server config.

```yaml
server:
  proxy_auth_header: X-Remote-User

indexer:
  languages:
    - en
    - de
  language_detection_accuracy: high
```

— Config keys and migration caveats both named. v0.21.0

07 **Configurable upload and extraction size limits** NEW

70

`server.max_batch_body_size` controls document, PDF, and batch upload limits, defaulting to 40 MiB; HTTP HTML downloads and extracted text now have a separate 32 MiB ceiling.

— Exact config key and both size limits given. v0.21.0

08 **MCP client request and pagination restrictions** BREAKING

70

MCP clients sending requests with an unrelated `Origin` are now rejected; set `server.base_url` to the public URL when running behind a reverse proxy. Clients paging through opened history must pass both `last_id` and `last_updated_at` from the previous response.

— Config key and required pagination fields named. v0.21.0

09 **Symlink and Org-file indexing safety fix** BREAKING

65

Previously indexed files reached through symbolic links outside configured directories are removed by running `hister reindex`; Org files using `#+INCLUDE` or `#+SETUPFILE` must now contain the needed text directly rather than pulling it in from elsewhere.

— Exact remediation command and affected directives named. v0.21.0

THINNER COVERAGE BELOW

10 Public version reporting and Prometheus metrics NEW

55

`server.expose_version` enables public version reporting for monitoring clients (disabled by default), and Prometheus latency metrics now use native histograms.

— Config key named; metric change only briefly noted. v0.21.0

└──▷ BREAKING ON UPGRADE

! Existing GitHub account links are not automatically migrated to permanent account IDs — back up the database and arrange verified account relinking before users sign in, or login will create a separate account while documents remain with the old one.

! The old `create-user`, `show-user`, `update-user`, and `delete-user` command names are deprecated aliases; account operations now target the running server and require an administrator's personal token. The old token display flag is replaced by `--show-token`; `--token` now supplies authentication credentials.

! Uppercase `AND`, `OR`, and `NOT` are now operators — quote them to search for the literal words. Negation takes precedence over `AND`, then `OR`; malformed queries are rejected by search, update, and delete operations.

! Run `hister reindex` to remove previously indexed files reached through symbolic links outside configured directories. Org files using `#+INCLUDE` or `#+SETUPFILE` must contain the needed text directly.

! Document, PDF, and batch uploads now default to a 40 MiB limit controlled by `server.max_batch_body_size`; HTTP HTML downloads and extracted text have a 32 MiB ceiling.

! MCP clients: requests carrying an unrelated `Origin` are rejected. Set `server.base_url` to the public URL when using a reverse proxy. Clients paging opened history must pass both `last_id` and `last_updated_at` from the previous response.

! Changing OIDC's `identity_claim` requires relinking existing accounts; leaving it unset preserves the existing email-based identity.

! Run `hister reindex` to apply language detection changes to existing documents; reapply language corrections made before persistent pins were available if they should survive future updates.

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Agentic Infrastructure

Google Gemini Enterprise Agent Platform

Feedback ⓘ changelog · Oct 28, 2026 CHANGELOG ↗

Gemini Enterprise Agent Platform provides Google's managed model training, inference, evaluation, and agent deployment services, formerly offered under Vertex AI.

Adds Cloud Storage image output config, per-policy intent-matching control, `cm health` pre-flight command, and GA supervised fine-tuning for `gemini-3.5-flash-lite` with CMEK.

Adds `sgpUserIntentMatching` key on authorization extension metadata to enable or disable user intent matching for semantic governance policies; set to `DISABLED` to reduce indirect prompt-injection protection, and check the mode used per evaluation in `sgpUserIntentMatchingMode` policy evaluation logs.

Adds `outputConfig.outputUriPrefix` and `generationConfig.responseFormat[].image.delivery` (value `"URI"`) to Gemini image generation requests, writing generated and edited images directly to a Cloud Storage bucket and returning the object URI in `fileData.fileUri` instead of inline bytes.

Adds `cm health` command (with `--strict` and `--timeout` flags) in CodeMender v0.16.0 to verify credentials, configuration, endpoint reachability, and project authorization before running scans.

UPDATED VULNERABILITY CHECK CodeMender v0.16.0 now runs a concurrent file-by-file deep scan by default for higher vulnerability recall; pass `--quick` flag (`cm find --quick`) to run a single-session quick scan instead.

Supervised fine-tuning for Gemini 3.5 Flash Lite (`gemini-3.5-flash-lite`) is now generally available, with customer-managed encryption keys (CMEK) supported for tuning job resources and tuned models.

+2 more

└──▷ BREAKING ON UPGRADE

! CodeMender now runs a deep scan by default instead of a quick scan; pipelines or scripts that relied on the previous default quick-scan behavior must pass `--quick` explicitly to preserve it.

WAS THIS USEFUL? 👍 👎

Hugging Face

Feedback ⓘ changelog · Oct 28, 2026 CHANGELOG ↗

RANK

Hugging Face is an open platform for building, sharing, and deploying machine-learning models, datasets, and applications.

> Hugging Face rolled out external collaborator invitations for private repos, new 3D asset preview tooling for OpenUSD and SimReady files, and passkey-based sign-in and MFA.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 External collaborator invites for private repos NEW

80

Organization admins can invite external collaborators directly from a private repository's settings page, granting read or write access scoped to that repository's Resource Group only. A new External invitations setting in Resource Groups (off by default, available on Team and Enterprise plans only) lets resource group admins invite external collaborators once an org admin has enabled it.

Enable external collaborator invitations for a resource group so group admins can onboard outside contributors to private repos without org-admin involvement each time.

> 📍 In the Hub, go to your Organization › Resource Groups › [Your Group] › Settings, and toggle External invitations to on.

– Names exact setting and scope, plus step-by-step toggle path

changelog-20261028-3439aa9e

02 3D OpenUSD and SimReady asset previews NEW

75

Adds an interactive 3D viewer for OpenUSD files (`.usd` , `.usda` , `.usdc` , `.usdz`) in datasets, models, and buckets, loading layers, payloads, and textures, alongside a 3D Assets preview panel on dataset pages for SimReady packages that shows the number of packages across the dataset. The viewer itself is published as open-source `@huggingface/usd-web-viewer` in the `huggingface.js` library.

– Names file formats and package but no exact UI steps changelog-20261028-3439aa9e

03 Passkey sign-in and two-factor authentication NEW

70

Supports passkeys (fingerprint, face, screen lock, or security key) as a primary sign-in method and as a second factor alongside an authenticator app, configurable from the Hub's Authentication settings.

Add a passkey as a second factor to harden account security with a hardware key or biometric instead of relying solely on a TOTP app.

> 📍 In the Hub, go to your profile › Settings › Authentication settings, then select Add a passkey and follow the browser prompt to register your fingerprint, face, screen lock, or security key.

— Exact navigation and registration steps provided `changelog-20261028-3439aa9e`

WAS THIS USEFUL? 👍 👎

HeyGen HyperFrames

Feedback ⓘ v0.8.210 NOTES ↗ RANK

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

> HyperFrames shipped a ProRes profile flag for MOV renders, expanded ffmpeg build diagnostics in `doctor`, and added seek-loop profiling, local HTML capture, and a new lint warning across two releases.

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

01 ProRes profile selection for render command NEW 80

The `render` command gains a `--prores-profile` flag accepting `4444`, `hq`, `422`, `lt`, or `proxy` profiles for MOV output, letting users target specific ProRes quality/size tradeoffs.

Render a project to a high-quality ProRes MOV for delivery to a post-production pipeline that requires lossless-grade codec quality.

```
$ hyperframes render --prores-profile hq output.mov
```

— Exact flag, values, and runnable command given v0.8.210

THINNER COVERAGE BELOW

02 ffmpeg build diagnostics in doctor IMPROVED 47

The `doctor` command now names the tested ffmpeg build and notes that loudness comparisons are only valid when run against that same build.

— Names command and behavior but no exact output format v0.8.210

03 Seek-loop profiling in check command NEW 47

The `check` command now reports where its seek loop spends its time, surfacing render bottlenecks to the user.

— Names command and purpose, no sample output shown v0.8.208

04 Lint warning for sub-composition root ID collisions NEW 40

Adds a `Lint` warning when a sub-composition's linked stylesheet names its root ID, flagging a potential conflict.

— Names the lint feature but no example or config shown v0.8.208

05 Local HTML file capture restored IMPROVED 30

The CLI restores capture support for local HTML files, which had previously been unsupported.

— Thin description, no flag or command named `v0.8.208`

WAS THIS USEFUL? 👍 👎

Tracer Cloud OpenSRE

Feedback ⓘ v0.1.2026.10.28 NOTES ↗

RANK

OpenSRE is an open-source toolkit for building AI SRE agents.

> OpenSRE simplified its CLI by combining provider login and model selection into a single flow.

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

01 Unified login and model selection CLI step IMPROVED

40

Provider login and model selection are now combined into a single CLI step, via the feat(cli) change (#6919).

— Names PR and CLI area but not exact command or flags `v0.1.2026.10.28`

WAS THIS USEFUL? 👍 👎

Snowflake Cortex AI

Feedback ⓘ

October 28, 2026: Cortex Agents Coding Agent thinking effort, system instructions, and s

NOTES ↗ RANK

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

> Snowflake Cortex AI's Coding Agent API gained three new request-level controls for reasoning depth, system prompt customization, and sandbox file access.

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

01 **Coding Agent API thinking effort, system instructions, and stage mounts**

NEW 88

The `agent:run` API now supports a `thinking_effort` field (when using `code_toolset_all`) accepting `minimal`, `low`, `medium`, `high`, `xhigh`, or `max` to control the orchestration model's reasoning effort, with requests above a model's supported subset silently capped. It also adds `instructions.system` to append custom system instructions to the managed coding-agent prompt without replacing sandbox, tool, or skill instructions, and `stage_mounts` on the `code_toolset_all` entry in `tool_resources` to mount a Snowflake stage or versioned stage (notebooks, Streamlit apps, workspaces, agents) into the coding sandbox.

Set maximum reasoning effort, inject custom system instructions, and mount a Snowflake stage into the coding sandbox in a single Coding Agent request.

```
$ curl -X POST
https://<account>.snowflakecomputing.com/api/v2/cortex/agent:run
\
  -H 'Authorization: Bearer <token>' \
  -H 'Content-Type: application/json' \
  -d '{
    "model": "<model-name>",
    "thinking_effort": "high",
    "instructions": {
      "system": "Always return results as a pandas DataFrame and
include row counts."
    },
    "tools": [{"type": "code_toolset_all"}],
    "tool_resources": {
      "code_toolset_all": {
        "stage_mounts": ["@MY_DB.MY_SCHEMA.MY_STAGE"]
      }
    },
    "messages": [{"role": "user", "content": "Analyse the CSV
files on the mounted stage and summarise trends."}]
  }'
```

— Names exact fields, values, and shows a runnable curl request.

October 28, 2026: Cortex Agents Coding …

WAS THIS USEFUL? 👍 👎

AWS Amazon Bedrock AgentCore

Feedback ⓘ

Runtime: Mount file systems with FUSE in your container (2026-10-27) BLOG ↗

RANK

Amazon Bedrock AgentCore provides managed agent runtimes, memory, identity, gateways, observability, and evaluation services.

> Amazon Bedrock AgentCore Runtime added support for mounting remote storage as local files inside agent containers.

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

01 **FUSE file system mounting in Runtime** NEW 45

AgentCore Runtime now supports mounting file systems with FUSE (Filesystem in Userspace) inside agent containers, letting agents work with remote storage as local files at runtime without orchestrating explicit downloads and uploads.

— Explains mechanism but no config key or command given

Runtime: Mount file systems with FUSE i…

WAS THIS USEFUL? 👍 👎

Triton

Feedback ⓘ v3.9.0 NOTES ↗ RANK

Triton is a Python-embedded language and compiler for writing GPU kernels, and is the backend most PyTorch and inference-engine kernels are generated through.

> Triton 3.9.0 adds new atomic primitives (`tl.atomic_load` , `tl.atomic_store` , `tl.atomic_poll`) and approximate FP32 division, expands its GSan and FpSan sanitizers to cover more execution modes, and ships several breaking changes to atomic and float-cast default semantics alongside notable compile-time and kernel-throughput speedups.

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

01 New tl.atomic_load, tl.atomic_store, tl.atomic_poll ops NEW

90

Adds `tl.atomic_load` and `tl.atomic_store` , defaulting to `acquire` and `release` semantics respectively and available in Gluon and the interpreter, plus `tl.atomic_poll` , which spins until a pointer's value equals an expected value and returns `False` on timeout.

Poll a shared counter until a peer CTA signals completion, with a timeout guard.

python

```
import triton
import triton.language as tl

@triton.jit
def wait_kernel(flag_ptr, out_ptr):
    # Returns False if the expected value is not seen within the
timeout
    timed_out = tl.atomic_poll(flag_ptr, 1)
    tl.store(out_ptr, tl.where(timed_out, -1, 0))
```

— Names exact ops and default semantics, with a runnable example. v3.9.0

02 Approximate FP32 division via tl.fdiv NEW

75

Adds tl.fdiv(..., approx=True) for fast approximate FP32 division, using a fast hardware reciprocal instead of full IEEE division.

Use approximate FP32 division for maximum throughput in a normalisation kernel where exact rounding is not required.

```python
import triton
import triton.language as tl

@triton.jit
def fast_normalize_kernel(x_ptr, scale_ptr, out_ptr, N:
tl.constexpr):
    offs = tl.arange(0, N)
    x = tl.load(x_ptr + offs)
    scale = tl.load(scale_ptr)
    # approx=True uses fast hardware reciprocal — avoids full
IEEE division
    out = tl.fdiv(x, scale, approx=True)
    tl.store(out_ptr + offs, out)
```

— Named flag with runnable example, mechanism lightly described. v3.9.0

---

03 **Relaxed semantics option for RMW and CAS atomics** BREAKING

75

Adds `sem='relaxed'` option on read-modify-write and CAS atomics to skip CTA-wide barrier insertion; the default atomic ordering is now `acq_rel`, a breaking change meaning kernels relying on relaxed ordering must now pass `sem='relaxed'` explicitly.

— Names flag and exact default change, no code example. v3.9.0

---

04 **FpSan narrowing cast truncation behaviour** BREAKING

70

FpSan narrowing casts now fold discarded bits by default, a breaking change from previous truncation behaviour; set `TRITON_FPSAN_HOMOMORPHIC_CASTS=1` to restore the previous truncation behaviour.

— Names env var and exact before/after default change. v3.9.0

---

05 **E8M0 zero and NaN decode values changed** BREAKING

60

E8M0 `0x00` now decodes to `2^-127` and `0xFF` to NaN (previously zero and infinity), changing results for kernels that use those values.

— Exact before/after values named, no migration example. v3.9.0

---

06 **GSan support for CUDA graphs and multi-CTA kernels** IMPROVED

60

Extends GSan to support programmatic dependent launch, CUDA graphs with explicit dependencies, and multi-CTA kernels, running up to 3.8x faster on MoE workloads.

— Names scope and perf number but no usage example. v3.9.0

07 **Expanded Python syntax support in kernels** NEW 60

Kernel syntax now supports chained comparisons (e.g. `0 <= x <= 8`), keyword-only parameters, tuple unpacking in list comprehensions, and operator methods on aggregate types.

— Lists all four syntax additions, no example provided. v3.9.0

THINNER COVERAGE BELOW

08 **FpSan support for clamp, inline assembly, Gluon ops** IMPROVED 45

Extends FpSan to support `tl.clamp`, pure inline assembly, and Gluon embedding ops.

— Lists supported ops but no mechanism or example. v3.9.0

09 **Integer tensor support in tl.clamp** IMPROVED 35

Extends `tl.clamp` to accept integer tensors.

— Bare one-line extension, no example or detail. v3.9.0

10 **Up to 40x speedup for batched MMAv2 dots** IMPROVED 30

Delivers up to 40x speedup for batched MMAv2 dots on NVIDIA hardware.

— Bare performance claim, no mechanism or benchmark detail. v3.9.0

11 **Faster compilation for deeply nested loops** IMPROVED 30

Delivers about 10x faster compile times for deeply nested loops at nesting depth 8.

— Single metric, no mechanism explained. v3.9.0

└──▷ **BREAKING ON UPGRADE**

! Read-modify-write and CAS atomics now default to `acq_rel` semantics, inserting CTA-wide barriers; kernels relying on relaxed ordering must now pass `sem='relaxed'` explicitly.

! E8M0 `0x00` now decodes to `2^-127` and `0xFF` to NaN (previously zero and infinity), changing results for kernels that use those values.

! FpSan narrowing casts now fold discarded bits by default; set `TRITON_FPSAN_HOMOMORPHIC_CASTS=1` to restore the previous truncation behaviour.

WAS THIS USEFUL?  

OpenAI Python SDK

Feedback ⓘ v3.40.0 NOTES ↗ RANK

The OpenAI Python SDK provides typed Python clients for OpenAI model APIs, streaming, and asynchronous requests.

> OpenAI Python SDK v3.40.0 adds TTL options for prompt cache breakpoints, giving finer control over cache lifetime.

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

01 **TTL options for prompt cache breakpoints** NEW 33

Adds prompt cache breakpoint TTL options to the API, enabling control over how long prompt cache breakpoints are retained.

— No parameter name, defaults, or usage example given v3.40.0

WAS THIS USEFUL? 👍 👎

◆ **Inference Providers & GPU Clouds**

AWS Amazon SageMaker AI

Feedback ⓘ

changelog · Oct 28, 2026 CHANGELOG ↗ RANK

Amazon SageMaker AI provides managed tools and infrastructure for building, training, fine-tuning, and deploying machine-learning models.

> Amazon SageMaker AI added support for sending OTLP metrics from HyperPod cluster instances to Amazon CloudWatch.

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

01 **OTLP metrics from HyperPod to CloudWatch** NEW 48

SageMaker HyperPod cluster instances can now send OpenTelemetry Protocol (OTLP) metrics to Amazon CloudWatch, enabled via updated HyperPod policies.

— Names mechanism and destination but no config steps shown changelog-20261028-f8029554

WAS THIS USEFUL? 👍 👎

CoreWeave

Feedback ⓘ snapshot-20261028 NOTES ↗ RANK

CoreWeave is an AI cloud platform offering GPU compute, managed Kubernetes and storage for training and inference.

> CoreWeave's SUNK v8.2.0 release adds gang-scheduling for Kubernetes Jobs and LeaderWorkerSets, switches topology configuration to a new file format, and ships several breaking default changes to login Pods alongside a Slurm upgrade.

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

01 **topology.yaml replaces topology.conf** BREAKING 65

SUNK now generates `topology.yaml` instead of `topology.conf`. This is a breaking change: any configuration or tooling that references `topology.conf` will break on upgrade.

— Exact file names given plus clear upgrade warning snapshot-20261028

THINNER COVERAGE BELOW

02 **Gang-scheduling for Jobs and LeaderWorkerSets** NEW 40

The SUNK Pod Scheduler now gang-schedules Kubernetes Jobs and LeaderWorkerSets, ensuring all pods in these workloads are scheduled together.

— Names the mechanism and two workload types but no further detail snapshot-20261028

03 **Login Pod shell and default behavior changes** BREAKING 40

Login Pods add `zsh` as an available shell, and per-user login Pods are now off by default, changing behaviour for clusters that previously relied on the old default.

— Names both changes but no migration steps given snapshot-20261028

04 **Slurm upgraded to 25.11.8-coreweave.2** IMPROVED 40

SUNK upgrades its bundled Slurm version to `25.11.8-coreweave.2`.

— Exact version named but no detail on what changed snapshot-20261028

05 **Unspecified default changes in SUNK v8.2.0** BREAKING 15

The release notes warn that several other unspecified defaults change in SUNK v8.2.0 and advise reviewing configuration before upgrading, without naming them.

— Vague warning with no named surfaces or specifics `snapshot-20261028`

└──▷ BREAKING ON UPGRADE

! `topology.conf` is replaced by `topology.yaml` — any configuration or tooling that references `topology.conf` will break on upgrade.

! Per-user login Pods are now off by default, changing behaviour for clusters that previously relied on the old default.

! Several other unspecified defaults change in SUNK v8.2.0 — review your configuration before upgrading.

WAS THIS USEFUL? 👍 👎

Eigen Labs Darkbloom

Feedback ⓘ v0.10.4 NOTES ↗

CODE ↗ RANK

Darkbloom is a for the innovators pushing the frontier.

> Darkbloom v0.10.4 introduces delta checkpoint writing for Qwen models with a reversion switch, forced tool_choice support for specific model families, and a more detailed `darkbloom status` output.

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

01 **Delta checkpoints for Qwen recurrent layouts** NEW 86

Darkbloom now writes delta (incremental) complete checkpoint files for Qwen native recurrent layouts, reducing checkpoint write overhead without requiring migration. The `DARKBLOOM_PREFIX_CACHE_SSD_INCREMENTAL_CHECKPOINTS=0` environment variable reverts to writing complete (non-delta) checkpoint files, is forwarded to the launchd provider job, and takes effect at the next `darkbloom start`.

— Names exact env var and mechanism, lacks performance numbers v0.10.4

02 **Forced tool_choice support for select models** NEW 67

Adds forced `tool_choice` support (`required` and named functions) for Qwen3.5-9B, Qwen3.5-35B-A3B, Qwen3.6-35B-A3B, and MiMo-V2.6-Flash, advertised via `tool_constraint_models`.

— Names models and field but no usage example v0.10.4

WAS THIS USEFUL? 👍 👎

Fireworks AI

Feedback ⓘ New model on Serverless: Rune 26B-A4B

NOTES ↗ RANK

Fireworks AI provides hosted inference, model fine-tuning, and deployment APIs for open-weight models.

> Fireworks AI added the Rune 26B-A4B decision model to Serverless, enabling document-grounded probability answers to large batches of typed questions at zero output-token cost.

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

01 **Rune 26B-A4B decision model on Serverless** NEW 86

Adds `accounts/fireworks/models/rune` to Serverless, callable via the System One decisions API or the OpenAI-compatible Decisions API, accepting text-only input up to 262K context. It supports up to 256 typed questions per document in multiple-choice, true/false, or graded-scale formats, returning a calibrated probability for each from a single read with zero output tokens (billed for input only).

— Names exact model id, endpoints, context limit, question formats, and billing.

New model on Serverless: Rune 26B-A4B

WAS THIS USEFUL? 👍 👎

Modal

Feedback ⓘ 1.6.3 (2026-10-27) NOTES ↗ RANK

Modal is a cloud infrastructure for teams that develop, train, and serve AI applications at scale.

> Modal 1.6.3 reworks how container images pull and cache base layers from private registries, adds a programmatic App.stop() method and Shared Endpoints, and gives workspace admins new CLI controls for member tokens and Image Builder versions.

└──▷ **WHAT SHIPPED · 7 FEATURES** most completely described first ? what's the number?

01 **Private registry pulls and layer caching in Image.from_dockerfile()**

IMPROVED 81

Adds a `registry_secret=` parameter to Image.from_dockerfile() to pull a private base image, limited to single-stage Dockerfiles. Also adds `force_build=True` support to invalidate the Dockerfile-keyed build cache, and caches the `FROM` instruction independently as a separate Modal Image layer so it can be reused across builds sharing the same base image.

– Named params and caching mechanism given, no full runnable example 1.6.3 (2026-10-27)

02 **Multi-cloud registry support in Image.from_registry()**

IMPROVED 79

Image.from_registry() now pulls images from AWS ECR or GCP Artifact Registry when `secret=` keys carry the appropriate credentials, superseding the separate Image.from_aws_ecr() and Image.from_gcp_artifact_registry() constructors. It also adds the Python distribution via `add_python=` in a separate layer so the Python version can be changed without invalidating the base layer cache.

– Mechanism and deprecated constructors named, lacks code example 1.6.3 (2026-10-27)

03 **Shared Endpoints via --shared flag** NEW 73

Adds a `--shared` flag to `modal endpoint create` to create Shared Endpoints, letting external consumers call them without authenticating to your Workspace.

Create a Shared Endpoint so external consumers can call it without authenticating to your Workspace.

```
$ modal endpoint create my-function --shared
```

– Runnable CLI command included 1.6.3 (2026-10-27)

04 **Workspace member token management** NEW 69

Adds a `Workspace.member_tokens` SDK namespace and a `modal workspace member-tokens` CLI command to list and delete tokens associated with Workspace members.

— Named SDK and CLI surfaces, no sample command shown 1.6.3 (2026-10-27)

05 **App.stop() method** NEW 66

Adds an App.stop() method to programmatically stop a running App, usable from a script or automation pipeline without visiting the dashboard.

Stop a long-running App from a script or automation pipeline without visiting the dashboard.

python

```python
import modal

app = modal.App.from_name('my-app')
app.stop()
```

— Comes with a runnable code snippet 1.6.3 (2026-10-27)

06 **Environment-level Image Builder version control** NEW 62

Adds a `--set-image-builder-version` flag to `modal environment update` to configure the Image Builder version at the Environment level.

— Exact flag named, no example run shown 1.6.3 (2026-10-27)

THINNER COVERAGE BELOW

07 **Image Builder version 2026.10** NEW 48

Introduces Image Builder version 2026.10 with updates to base Debian images, Python versions, and Modal dependencies bundled in Function containers.

— Version named but no detail on how to opt in beyond the flag 1.6.3 (2026-10-27)

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

llmfit

Feedback ⓘ v1.1.18 NOTES ↗ CODE ↗ RANK

Hundreds of models & providers. One command to find what runs on your hardware.

> llmfit v1.1.18 adds LM Studio benchmarking support, visibility into backend-incompatible models, NVIDIA PCIe link detection, and brings the web dashboard to feature parity with the TUI.

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

01 **Web dashboard parity with TUI filters** IMPROVED 55

The web dashboard now includes estimate confidence, concurrency, storage, and TUI filters, bringing it to parity with the terminal interface.

— Lists specific added dashboard capabilities, no exact UI path v1.1.18

02 **LM Studio model benchmarking** NEW 45

The bench subsystem can now discover and benchmark models that are already loaded in LM Studio directly, rather than requiring a separate benchmarking flow.

— Names LM Studio and bench subsystem but no exact command shown v1.1.18

03 **NVIDIA PCIe link detection** NEW 40

Hardware telemetry now detects and displays the PCIe link of NVIDIA GPUs.

— Names the hardware signal detected but no usage path given v1.1.18

04 **CLI flag to show backend-incompatible models** NEW 35

A new CLI option lets users request that backend-incompatible models be shown, rather than hidden by default.

— Mentions a new option but not its exact flag name v1.1.18

WAS THIS USEFUL? 👍 👎

Jan AI Jan

Feedback ⓘ v0.8.8 NOTES ↗ RANK

Jan runs language models locally and provides a desktop chat interface.

> Jan v0.8.8 locks down its local API server with new Trusted Hosts and origin-checking, upgrades the bundled llama.cpp engine to 0.7.0 with new speculative-decoding settings, and tightens several other security defaults including CORS enforcement, MCP tool execution control, and restricted environment variables for Claude Code.

└──▷ **WHAT SHIPPED · 7 FEATURES** most completely described first ? what's the number?

01 Trusted Hosts and origin checking for local API server NEW

88

Adds Trusted Hosts to `Settings > Local API Server` so browser-based clients (Open WebUI, Obsidian plugins via `app://obsidian.md`, local files via `file://`, sandboxed frames via `null`) can be explicitly allowed to call Jan's local API. Pairs this with origin-checking so only same-machine, Jan-internal, or Trusted Hosts browser origins are accepted; SDK, script, and `curl` requests with no browser origin continue to work unchanged.

Allow an Open WebUI instance running on another machine to call Jan's local API without being blocked by origin checks.

> 📍 In Jan, go to Settings › Local API Server › Trusted Hosts and add the hostname or origin of your Open WebUI instance (e.g. `http://192.168.1.50:3000`).

Allow an Obsidian plugin to call Jan's local API by adding its app origin to Trusted Hosts.

> 📍 In Jan, go to Settings › Local API Server › Trusted Hosts and add `app://obsidian.md`.

— Names exact setting path and shows two worked examples v0.8.8

02 llama.cpp upgraded to 0.7.0 with speculative decoding settings

IMPROVED

80

Upgrades the bundled llama.cpp engine from 0.5.0 to 0.7.0 (b11669), adding three new per-model settings: `MoE expert GPU cache` for mixture-of-experts models, Probabilistic draft sampling, and Draft batch size for speculative decoding. Saved prompt caches from the 0.5.0 engine are discarded once on first run and rebuilt by the new engine.

Enable speculative decoding with the new llama.cpp 0.7.0 per-model settings for faster inference on a supported model.

> 📍 In Jan, open your model's settings and set Probabilistic draft sampling and Draft batch size to your desired values under the model configuration panel.

— Names exact version, build, and three new settings with example v0.8.8

---

03 **Restricted custom environment variables for Claude Code** 68

BREAKING

Restricts Custom environment variables for Claude Code integration — keys such as `PATH`, `PATHEXT`, `SystemRoot`, `ComSpec`, `LD_*`, `DYLD_*`, `JAN_*`, and GPU runtime paths are now refused with a message, while other variables including `NODE_OPTIONS` continue to save as before.

— Names every refused key but no reconfiguration example v0.8.8

THINNER COVERAGE BELOW

04 **CORS switch now enforced on local API server** BREAKING 55

Enforces the `CORS` switch in `Settings > Local API Server`, which was previously shown but had no effect; it now starts `on` after updating, meaning browser-based clients that relied on it being silently ignored will be blocked unless trusted.

— Names setting and default change, no walkthrough given v0.8.8

05 **Per-endpoint MCP tool execution control** NEW 45

Controls server-side MCP tool execution per-endpoint via `Settings > Local API Server`, keeping it opt-in for every API request.

— Thin one-line description, no mechanism or example v0.8.8

06 **Links always open in system browser** IMPROVED 40

Links in model replies, tool results, model cards, and release notes now always open in the system browser rather than inside Jan's window; non-web/non-email links render as plain text instead of becoming clickable.

— Describes behavior change only, no setting or flag v0.8.8

07 **MCP sign-in endpoint validation** BREAKING 35

MCP servers whose sign-in endpoints are not `http` / `https` addresses now show a sign-in error instead of proceeding with sign-in.

— Minimal detail, no example or workaround given v0.8.8

└──▷ **BREAKING ON UPGRADE**

! The `CORS` switch in `Settings > Local API Server` now takes effect and defaults to `on` after updating — browser-based clients that previously relied on it being silently ignored will now be blocked unless their origin is trusted.

! Browser-based clients on remote hosts (e.g. Open WebUI direct connection, browser extensions) must have their host added to `Settings > Local API Server > Trusted Hosts` or requests will be rejected.

! MCP servers whose sign-in endpoints are not `http` / `https` addresses now show a sign-in error instead of proceeding.

! Saved prompt caches from llama.cpp 0.5.0 are discarded once on first run and rebuilt by the new 0.7.0 engine.

! Custom environment variables for Claude Code integration that match `PATH`, `PATHEXT`, `SystemRoot`, `ComSpec`, `LD_*`, `DYLD_*`, `JAN_*`, or GPU runtime paths are now refused and will not be saved.

WAS THIS USEFUL? 👍 👎

vMLX

Feedback ⓘ v1.6.82 NOTES ↗

The vMLX server runs compressed MLX models on Apple Silicon with disk caching, paged memory, continuous batching, and hybrid SSM scheduling.

vMLX 1.6.82 ships DeepSeek-V4.1 with DSpark adaptive speculative decode, disk-only SSD prefix-cache reuse, multimodal VL serving, native DSML tool calls, and new JANGT sub-2-bit trellis and Metal MoE decode kernels.

└──▷ GET THIS VERSION ▾

pipx

```
$ pipx install vmlx==1.6.82
# already installed? upgrade to this version:
$ pipx install --force vmlx==1.6.82
```

git

```
$ git clone --branch v1.6.82 https://github.com/jjang-ai/vmlx.git
# already have the repo? check out this version:
$ git checkout v1.6.82
```

PyPI attestation · checked 28 Oct Wrong command?

NEW SUPPORTED PLATFORM Adds `deepseek_v41` model family (`DeepSeek-V4.1-Flash`) with DSpark adaptive speculative decode enabled by default, targeting 50+ tok/s on code (path `dsv41_dspark`).

Adds disk-only SSD prompt-boundary cache reuse for DeepSeek-V4.1 (DSV4-parity), with bit-exact 4-bit prompt snapshots and typed disk restore.

Adds in-app multimodal (image) serving for DeepSeek-V4.1 via text-routed VL path, including image serving via `/v1/responses` mirroring `/v1/chat/completions` .

Adds native V4.1 DSML tool-call support (parser + prompt render + fallback detection) with V4.1-only tool-parser encoder.

Adds `dsa_pool_index_scores` / `dsa_index_scores_mpp` Metal kernel for DSA pool index scoring on M5 matrix units for long-context prefill.

+7 more

WAS THIS USEFUL? 👍 👎

Ollama

Feedback ⓘ v0.40.12 NOTES ↗ CODE ↗

Ollama downloads, serves, and chats with local language models through a CLI and REST API.

Ollama v0.40.12 adds Liquid AI LFM3 and d1 2B model families, DFlash2 speculative decoding on MLX, and faster Apple silicon prompt processing.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.40.12 https://github.com/ollama/ollama.git
# already have the repo? check out this version:
$ git checkout v0.40.12
```

NEW SUPPORTED PLATFORM Adds `lfm3` (1.7B, 256K-token context), `lfm3-vl` (image understanding), `lfm3-audio` (speech in 18 languages), and `lfm3-omni` (text, images, and audio) via `ollama run lfm3`.

Adds `d1:2b` and `d1-omni` decision models built on LFM3 via `ollama run d1:2b`.

Speculative decoding now supports DFlash2 draft models on MLX, up to 1.25x generation throughput improvement over MTP.

Faster prompt processing on Apple silicon for Gemma 4, Flint, and Laguna models — up to +27% prefill throughput on Apple M5 Max.

UPDATED INTEGRATION OpenCode v2 integration

WAS THIS USEFUL? 👍 👎

◆ AI / LLM Gateways

Mozilla AI Otari

Feedback ⓘ v0.33.0 NOTES ↗ RANK

Open-source, OpenAI-compatible LLM gateway you run yourself. One endpoint for 40+ providers, with virtual keys, budgets, and usage tracking.

> Otari v0.33.0 adds geo-aware provider routing with a UI-managed country file, expands agent trace views with context-window and session-cost detail, and brings Gemini models into the agent gateway, alongside two breaking changes to search/fetch endpoints and catalog update delivery.

└──▷ **WHAT SHIPPED · 9 FEATURES** most completely described first ? what's the number?

01 Geo-aware provider routing NEW

80

Otari now looks up the country a request originates from and skips providers that do not serve that country, with the gate applied across chat, files, batches, and all other provider routes. Operators define which countries each hosted provider serves via a JSON country file that can be uploaded from the settings page and is stored in the database.

Upload a country-to-provider mapping so Otari can gate requests to providers by the caller's origin country.

> 📍 In the web UI, go to Settings › (country file upload section) and upload your JSON country file to populate the database.

Configure which countries a hosted provider serves so geo-aware routing can skip it for out-of-region requests.

> 📍 In the web UI, go to the hosted provider's settings page and set the countries it serves — sourced from the uploaded JSON country file.

— Names config surface and upload path, scope across routes v0.33.0

02 Agent trace view improvements IMPROVED

60

Trace views now show what fills a model request's context window, chart sessions active over time in buckets a session can fill, price a coding agent's session along with its turns and tool calls, and animate turns a live session adds into its trace tree — making a coding agent's session readable enough to debug in traces.

— Mechanism described but no concrete navigation or command v0.33.0

03 Search and fetch endpoints restricted to public addresses

60

BREAKING

An organization's search and fetch endpoint is now restricted to public addresses only — internal/private endpoints that were previously accepted will be rejected.

— Clear before/after behaviour for migration v0.33.0

THINNER COVERAGE BELOW

04 **Combined hosted-provider catalog updates** BREAKING 55

Hosted-provider catalog updates for prices and model lists are now delivered as a single combined update; any tooling that consumed these as separate updates must be adapted.

— States change and required adaptation, no exact schema given v0.33.0

05 **Rate-limit reporting mode** NEW 45

A new rate-limit reporting mode counts a rule per budget and can report a limit without enforcing it.

— Explains behaviour but no config key or flag given v0.33.0

06 **Web search dashboard page** NEW 45

Adds a Web search page in the dashboard to manage search and fetch tools.

Inspect search and fetch tool configuration from one central location.

📍 In the dashboard, navigate to the Web search page to manage search and fetch tools.

— Clear UI navigation given, limited mechanism detail v0.33.0

07 **Named extra request fields in chat completions** NEW 40

Provider instances can now accept named extra request fields in chat completions.

— Names surface (chat completions) but no field names given v0.33.0

08 **Gemini gateway parity for agents** NEW 35

Adds Gemini parity for gateway-routed agents, enabling Gemini models to be used through Otari's agent gateway.

— Names the integration but no mechanism or steps v0.33.0

09 **Per-attempt metrics for upstream requests** NEW 30

Otari now reports per-attempt metrics: each upstream attempt and every coded refusal.

— Bare description of what is reported, no usage path v0.33.0

└──▷ **BREAKING ON UPGRADE**

! An organization's search and fetch endpoint is now restricted to public addresses only — internal/private endpoints that were previously accepted will be rejected.

! Hosted-provider catalog updates for prices and model lists are now delivered as a single combined update; any tooling that consumed these as separate updates must be adapted.

WAS THIS USEFUL? 👍 👎

dario

Feedback ⓘ v6.18.1 NOTES ↗ RANK

One local endpoint that routes Cursor, Cline, Aider, Claude Code and the Agent SDK through your Claude and ChatGPT subscriptions, with failover and pooling and zero runtime dependencies.

> dario v6.18.1 restores ChatGPT prompt-cache affinity through new routing headers and fixes a gap in pool-fallback analytics reporting.

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

01 Prompt-cache affinity headers for ChatGPT requests IMPROVED

78

dario now adds `session-id`, `thread-id`, and `x-codex-routing-hint: model=<model>` headers to every ChatGPT (Codex) request so turns are routed to the machine holding the conversation's prompt-cache prefix, restoring cache affinity for both wire shapes and the raw `/v1/responses` path.

— Names exact headers and path but no setup steps v6.18.1

02 Pool-fallback requests counted in analytics IMPROVED

70

`--pool-fallback` requests served by a Codex seat are now counted in `/analytics` rows, `--log-file` lines, and ledger entries, filed under the substituted model and request number, with cache trace included.

— Names flag and reporting surfaces, fixes a gap v6.18.1

WAS THIS USEFUL? 👍 👎

◆ Inference Serving & Optimization

Niko1221 Strata

Feedback ⓘ v0.2.7 NOTES ↗ RANK

Strata is a local inference engine with one-click install for Windows and Linux, providing an OpenAI/Anthropic-compatible API on localhost for models like Qwen and DeepSeek.

> Strata v0.2.7 ships a dense batch of NVIDIA- and DeepSeek-specific performance tuning controls (expert routing, prefill layout, batch scheduling, kernel defaults), adds a DeepSeek vision tower, extends context windows to 512K/1M, and overhauls the web chat UI with French localization.

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

01 **NVIDIA expert-routing speed toggle** NEW 93

`STRATA_ROUTE_PRIOR=0` env var (or the 'Faster answers on 12 GB NVIDIA cards' toggle in Models > Native engine settings > Speed) controls residency-biased expert routing on discrete NVIDIA cards of 13 GiB or less, delivering up to +37% decode speed on an RTX 3060 12 GB.

— Exact env var, UI toggle path, and measured gain given. v0.2.7

02 **Layer-major prefill for dual-GPU splits** NEW 87

`STRATA_PREFILL_LAYER_MAJOR=0` env var controls layer-major prefill on two-CUDA-card splits, cutting time-to-first-token on a 32K prompt from 141.8 s to 27.7 s on two RTX A4000 cards.

— Named env var with precise before/after timing numbers. v0.2.7

03 **DeepSeek vision tower opt-out** NEW 81

`--vision off` / `--vision no` flags for `strata launch` and setup opt out of DeepSeek-V4.1-Flash's built-in vision tower (319 MiB, ~1 GB one-time download), which is now enabled by default.

Launch DeepSeek on a Ryzen AI Max (Strix Halo) Windows PC without the vision tower to save ~1 GB of download and RAM.

```
$ strata launch --vision off
```

— Exact flags, size figures, and a runnable example. v0.2.7

04 **Pipelined multi-card batch scheduling default** IMPROVED 79

`STRATA_BATCH_GROUPS_NV_SERIAL=1` env var reverts to the old serial batch scheduling; the new default is pipelined multi-card batching, giving +17–22% throughput at two-to-four concurrent chats on dual NVIDIA cards.

— New default explained with throughput numbers and opt-out var. v0.2.7

05 **DeepSeek RAM-store pre-fill toggle** IMPROVED 79

`DS4_SETTLE_STORE=1` env var restores DeepSeek's RAM-store pre-fill behavior, which is now off by default in exchange for faster prompt processing — up to +47% on an RTX 5070.

— Named env var and measured prompt-processing gain. v0.2.7

06 **DeepSeek expert-count quality flag** IMPROVED 73

`--experts 6` launch flag runs DeepSeek with six experts per token (Full quality mode); the new default is 5 experts (Faster mode).

Run DeepSeek in Full quality mode (six experts per token) instead of the new Faster default of five.

```
$ strata launch --experts 6
```

— Exact flag and usage example given, limited mechanism detail. v0.2.7

07 **GPU-specific kernel tuning defaults** NEW 62

New measured kernel speed tables are now the default for RTX 40-series (sm_89) cards, disabled via `STRATA_MMVQ_IL_89=0` ; new A100-class (sm_80) dequantisation and matrix-product tuning defaults can be disabled via `STRATA_DEQUANT_WIDE=0` and `STRATA_MMVQ_IL_SM80=0` .

— Opt-out vars named but no measured impact given. v0.2.7

08 **One-step defaults rollback env vars** NEW 60

`DS4_DEFAULTS=026` and `STRATA_DEFAULTS=026` env vars restore all 0.2.6 defaults for DeepSeek and Qwen respectively in a single step, for reverting the other 0.2.7 tuning changes at once.

Roll back all Qwen 0.2.7 behaviour changes in one step when troubleshooting a regression on a two-NVIDIA-card setup.

```
$ STRATA_DEFAULTS=026 strata launch
```

— Exact rollback vars with a runnable example, little mechanism detail. v0.2.7

09 **Extended context sizes for DeepSeek** NEW 60

512K and 1M context sizes are now selectable for DeepSeek models in Setup, Terminal, and the app, though marked not tested yet past 256K; the default remains 8K.

— Named sizes and surfaces but no command given. v0.2.7

THINNER COVERAGE BELOW

10 **Web chat UI overhaul** IMPROVED 8

Strata v0.2.7 includes a major overhaul of the web chat interface.

— Only mentioned in the release summary, no detail given. v0.2.7

11 **French localization** NEW 8

Strata v0.2.7 adds French localization.

— Only mentioned in the release summary, no detail given. v0.2.7

WAS THIS USEFUL? 👍 👎

compressed-tensors

Feedback ⓘ 0.20.0 NOTES ↗ RANK

Extends safetensors to store and load quantized, sparse, and otherwise compressed model weights.

> compressed-tensors 0.20.0 adds MXFP6 and MXFP4xFP8 quantization formats, non-uniform MoE expert sparsity support, and scoped layerwise decompress/recompress controls, alongside faster disk onloading, a fused NVFP4 kernel, and faster 4-bit unpacking, plus breaking changes including a Python 3.11 minimum and a new checkpoint-conversion worker default.

└──▷ **WHAT SHIPPED · 12 FEATURES** most completely described first ? what's the number?

01 MXFP6 and MXFP4xFP8 quantization formats NEW 90

Adds the `mxfp6-pack-quantized` format with `MXFP6A16` and `MXFP6A8` presets for 6-bit floating-point (E2M3) quantization, compatible with Quark/OCP packing and existing vLLM kernels, plus an `MXFP4xFP8` preset enabling MXFP4 weights paired with DeepSeek-style FP8 block-quantized activations.

Quantize all Linear layers of a model to MXFP6 (E2M3) and save in the `mxfp6-pack-quantized` format for direct vLLM inference.

python

```
from compressed_tensors.quantization import QuantizationConfig,
apply_quantization_config
from compressed_tensors.compressors import ModelCompressor
from transformers import AutoModelForCausalLM
import torch

model = AutoModelForCausalLM.from_pretrained('meta-llama/Meta-
Llama-3-8B', torch_dtype='auto', device_map='cuda:0')

config = QuantizationConfig(
    config_groups={'MXFP6A16': ['Linear']},
    ignore=['lm_head'],
)
apply_quantization_config(model, config)

compressor = ModelCompressor.from_pretrained_model(model)
compressor.compress_model(model)
model.save_pretrained('./Meta-Llama-3-8B-MXFP6')
compressor.update_config('./Meta-Llama-3-8B-MXFP6')
```

— Full format and preset names with a runnable quantize-and-save example. 0.20.0

02 **Scoped quantization via allowed_modules and leave_decompressed**

NEW

85

Extends `apply_quantization_config` with an `allowed_modules` argument to restrict quantization to specific modules, and a `leave_decompressed=False` option that fully removes quantization metadata after compression, enabling per-layer decompress/recompress pipelines.

Run layerwise decompress/recompress by restricting `apply_quantization_config` to specific modules and stripping metadata with `leave_decompressed=False`.

python

```python
from compressed_tensors.quantization import QuantizationConfig,
apply_quantization_config

config = QuantizationConfig(
    config_groups={'MXFP4': ['Linear']},
    ignore=['lm_head'],
)

# Only apply to a single layer; strip quantization metadata after
compression
apply_quantization_config(
    model,
    config,
    allowed_modules=['model.layers.0'],
    leave_decompressed=False,
)
```

— Named flags paired with a runnable layerwise decompress/recompress example. 0.20.0

03 **Auto-sized worker pool for checkpoint conversion** BREAKING

75

Changes `convert_checkpoint`'s default to `max_workers="auto"`, sizing the worker pool from shard count, per-job memory, free accelerator memory, and available CPUs; pass an explicit integer to restore the previous fixed-pool behaviour.

— Names the exact flag, its new default, and a migration path. 0.20.0

04 **Faster disk shard loading via handle reuse** IMPROVED

70

Extends `disk_load_context` to reuse safetensors handles and group subgraph reads, delivering 2.8×–6.3× faster shard reads (e.g. a Qwen3-MoE layer read drops from 11.0 ms to 4.4 ms).

– Concrete before/after benchmark but no direct usage snippet. 0.20.0

05 **Flexible weight staging for disk offload** BREAKING 65

Adds `OffloadCache.stage`, `DeviceCache` staging, memory pinning, and `device_map="disk"` support for flexible weight staging before GPU onloading. Custom `OffloadCache` subclasses must now implement stage() for subgraph staging to remain compatible.

– Names the new staging APIs and the breaking subclass requirement they introduce. 0.20.0

06 **Non-uniform MoE expert sparsity config** NEW 60

Introduces `ExpertSparsityConfig` for MoE non-uniform expert sparsity, supporting REAP (prune) and MoNe (novice) checkpoints; `MagnitudeExpertPruner` now also prunes per-expert router bias vectors.

– Names config class and pruning methods but has no usage example. 0.20.0

07 **Faster 4-bit weight unpacking** IMPROVED 60

Delivers 4-bit unpacking by splitting bytes instead of gathering words, cutting Llama-3.1-8B W4A16 CPU load time from ~11.5 s to ~3.8 s.

– Concrete before/after benchmark numbers but no runnable example. 0.20.0

THINNER COVERAGE BELOW

08 **Auto-discovery for FP8 block dequantization** NEW 50

Adds `FP8BlockDequantizer.from_pretrained` to auto-discover `weight_scale_inv` targets in a checkpoint's weight map.

– Names the method and target field but gives no example. 0.20.0

09 **Batched distributed offload metadata and shared tied-weight storage**

IMPROVED 45

Batches distributed offload metadata into one broadcast and barrier per dispatch instead of per tensor, and tied weights now share a single storage allocation on every rank.

– Describes the mechanism change but gives no numbers or usage. 0.20.0

10 **Fused NVFP4 quantize-and-pack Triton kernel** NEW 45

Delivers a fused NVFP4 quantize-and-pack Triton kernel that runs 1.07×–1.77× faster per weight.

– Gives a speedup range but no usage detail or flag name. 0.20.0

11 **QuantizationArgs default field serialization change** BREAKING

40

`QuantizationArgs` no longer serializes implicit default fields; code relying on those fields being present in serialized output must handle their absence.

— Names the affected class and behavior change with no migration example. `0.20.0`

---

12 **Python 3.11 minimum requirement** BREAKING

35

Python 3.11 or newer is now required; Python 3.10 and earlier are no longer supported.

— States the exact version cutoff with no further detail. `0.20.0`

└──▷ **BREAKING ON UPGRADE**

! Python 3.11 or newer is now required; Python 3.10 and earlier are no longer supported.

! `convert_checkpoint` now defaults to `max_workers="auto"`; pass an explicit integer to restore the previous fixed-pool behaviour.

! `QuantizationArgs` no longer serializes implicit default fields; any code that relied on those fields being present in serialized output will need to handle their absence.

! Custom `OffloadCache` subclasses must now implement stage() for subgraph staging.

WAS THIS USEFUL?  

Experiential Labs Experiential

Feedback ⓘ

v0.7.221 NOTES ↗ RANK

Experiential is the open source, zero markup gateway for BYOK, self-hosted and 1000+ marketplace models. It learns from your traffic to cut costs, recommend better models, and train a specialized model you own.

> Experiential v0.7.221 expands image generation support with new Azure and Bedrock Stability integrations plus a dedicated image-editing endpoint, and sharpens refusal diagnostics for image inputs.

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

01 **Image model support for Azure and Bedrock Stability** NEW

60

Adds Azure image model support and a Bedrock Stability integration, along with a new `/v1/images/edits` endpoint delivered via the Bedrock Stability integration for image editing requests.

– Names endpoint and providers but no usage detail or limits. v0.7.221

THINNER COVERAGE BELOW

02 **Clearer refusal diagnostics for image inputs** IMPROVED

30

Image-input refusals now name the carrying message for clearer diagnostics, and ledger entries now name the refused capability.

– Describes the change but no concrete surface or example. v0.7.221

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

rag-anything

Feedback ⓘ v1.4.4 NOTES ↗ RANK

RAG-Anything processes PDFs, images, tables, equations, audio, and video into a knowledge graph for intelligent multimodal RAG retrieval.

> RAG-Anything v1.4.4 locks down WeasyPrint-based Markdown-to-PDF conversion against remote and out-of-directory resources, adds a Docling-based parser path for Python 3.14 users who can't install MinerU 3.x, and ships several breaking changes to cache keys, output paths, and config access.

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

01 WeasyPrint local-only resource restriction and method selector

88

BREAKING

The WeasyPrint backend of `raganything.enhanced_markdown` now refuses remote resources and files outside the Markdown file's directory, permitting only `data:` URIs and local resources (CSS, fonts, images) within that directory; pass `base_url` to allow local resources in string-conversion calls. Markdown-to-PDF conversion gains an explicit `method="weasyprint"` selector to opt into this policy, since `method="auto"` prefers Pandoc when installed. PDFs previously generated from untrusted Markdown under earlier WeasyPrint configurations may have embedded referenced local files.

Convert untrusted Markdown to PDF with WeasyPrint's new local-only resource restrictions instead of Pandoc.

python

```python
from raganything.enhanced_markdown import
EnhancedMarkdownConverter, MarkdownConfig

converter = EnhancedMarkdownConverter(MarkdownConfig())
converter.convert_file_to_pdf(
    input_path="document.md",
    output_path="output.pdf",
    method="weasyprint",  # explicit; 'auto' would pick Pandoc if
installed
)
```

— Names exact module, flags, and security behaviour with runnable example v1.4.4

02 Office-conversion PDF output path changed

BREAKING

61

Office-conversion PDFs are now written to `<output_dir>/<stem>_<path-hash>/<stem>.pdf` instead of the previous path; any tooling reading from the old path will break.

— Exact new path format given, but no migration example v1.4.4

---

03 **MinerU version check tightening** IMPROVED

60

Tightens MinerU version checks; if an environment accidentally upgrades to MinerU 4, it can be recovered by reinstalling raganything, uninstalling `docvortex` and `mineru-llama-cpp`, then verifying with `mineru --version` and `pip check`.

Recover a Python 3.10–3.13 environment that accidentally upgraded to MinerU 4.

```
$ pip install -U raganything
pip uninstall -y docvortex mineru-llama-cpp
mineru --version
pip check
```

— Thin on mechanism but gives concrete recovery commands v1.4.4

THINNER COVERAGE BELOW

04 **Parse-cache keys now reflect VLM and OCR settings** IMPROVED

52

Includes `vlm_url` (hashed), PaddleOCR `cls`, and result-affecting Docling options in parse-cache keys, so cache entries correctly reflect those settings. Changes to these newly recognized options can invalidate existing parse caches on upgrade.

— Names exact fields but no steps to act on it v1.4.4

---

05 **Docling parser path for Python 3.14** NEW

50

Adds `parser="docling"` as the supported parsing path for Python 3.14 users, where MinerU 3.x cannot be installed.

— Names flag and reason but no usage example v1.4.4

---

06 **RAGAnythingConfig must be read from an instance** BREAKING

34

`RAGAnythingConfig` fields should now be read from an instance rather than from the class; class-level reads may no longer reflect environment defaults.

— States the change but no code example of fix v1.4.4

└──▷ **BREAKING ON UPGRADE**

! The WeasyPrint backend now refuses remote resources and files outside the Markdown file's directory; PDFs previously generated from untrusted Markdown with earlier WeasyPrint configurations may have embedded referenced local files.

! Office-conversion PDFs are now written to `<output_dir>/<stem>_<path-hash>/<stem>.pdf`; any tooling that read from the previous path will break.

! Changes to newly recognized parser options (`vlm_url`, PaddleOCR `cls`, Docling options) can invalidate existing parse caches on upgrade.

! `RAGAnythingConfig` fields should be read from an instance, not from the class; class-level reads may no longer reflect environment defaults.

WAS THIS USEFUL? 👍 👎

Weaviate

Feedback ⓘ v1.38.23 NOTES ↗ RANK

Weaviate is an open-source vector database that stores both objects and vectors, allowing for the combination of vector search with structured filtering with the fault tolerance and scalability of a cloud-native database.

> Weaviate's v1.38.23 release adds finer-grained observability to the text2vec-google module.

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

01 **Request duration metrics for text2vec-google module** IMPROVED

50

The `text2vec-google` module now exposes request duration metrics broken down by service, location, and model, allowing more granular monitoring of embedding request performance.

— Names metric breakdown but no metric names or endpoints v1.38.23

WAS THIS USEFUL? 👍 👎

ralforion OrionBelt Ontology Builder

Feedback ⓘ

v1.30.0 NOTES ↗ RANK

Browser-based ontology workbench for OWL ontologies and SKOS vocabularies. Streamlit + rdflib, no Java, no Protégé.

> OrionBelt v1.30.0 focuses on graph visualization and usability: SPARQL results can now be rendered directly on the graph, the path finder surfaces multiple routes, properties get custom edge colours, and node filters can be saved to file.

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

01 **Show on graph for SPARQL results** NEW 80

A Show on graph button appears under SELECT, CONSTRUCT, and DESCRIBE SPARQL query results, ringing matched entities in cyan and highlighting links between entities named in the same result row, with a count line and a Clear button to dismiss the highlight.

Visualise a SPARQL SELECT result directly on the graph to see which entities and relationships a query touches.

📍 In the SPARQL editor, run a SELECT query, then click Show on graph under the results table. Matched entities appear ringed in cyan; links between entities named in the same row are ringed too. Use the Clear button above the graph to dismiss the highlight.

— Names exact button, query types and reset control v1.30.0

02 **Custom edge colours per property** NEW 80

A Colours popover in Display options lets users assign a custom colour to a property's edges — covering domain/range edges, restrictions, assertions, and custom relations — with coloured properties named in the legend; the colour is kept with graph preferences rather than written into the ontology.

Assign a distinct colour to a heavily-used property so its edges stand out across the whole graph without touching the ontology file.

📍 Open Display options and click the Colours popover. Pick the property you want to highlight — the graph zooms to its edges and applies the chosen colour to every domain/range edge, restriction, assertion, and custom relation for that property. The colour is saved with your graph settings.

— Names popover, edge types and storage mechanism v1.30.0

03 **Adjustable text size setting** NEW 70

A Text size sidebar setting (90%, 100%, 115%, or 130%) scales app text and controls and persists across sessions, serving as the primary way to scale the desktop app without browser zoom.

— Exact values and location named, simple toggle v1.30.0

04 **Multi-route path finder with route picker** NEW 65

The path finder now surfaces up to 5 next-best routes, shortest first, with a Route picker to choose which is drawn and hop counts shown for equally short routes. In focus mode, path entities now stay on the graph regardless of the active focus, and focus seeds are left unchanged.

— Names picker and route count but no exact UI steps v1.30.0

05 **JSON import/export for node filter options** NEW 65

A File button beside `Paste / copy` in the Node options panel lets users download or load node filter/focus options as a JSON file named after the ontology, making complex filter configurations reusable across sessions.

Back up and restore a complex node-filter configuration so it can be reused across ontology sessions.

📍 Click the File button beside Paste / copy in the Node options panel to download the current filter as a JSON file named after the ontology. To restore it in another session, click File again and load the saved JSON.

— Named button and file format, clear usage path v1.30.0

THINNER COVERAGE BELOW

06 **Fullscreen graph view controls** IMPROVED 50

The fullscreen graph view gains Undo, Redo, search, and language pack controls placed in the gap before `Render`, since fullscreen mode hides the sidebar that normally hosts them.

— Names controls and placement but limited detail v1.30.0

WAS THIS USEFUL? 👍 👎

knowhere

Feedback ⓘ v1.2.34 NOTES ↗ RANK

Knowhere extracts, parses, and outputs structured chunks ready for AI Agents and RAG.

> Knowhere's latest releases focus on a progressive ingestion and retrieval pipeline that keeps a searchable baseline live while enhancements finish, plus page-memory table-of-contents ownership and citation identity preservation across the enhancement process.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 **Progressive ingestion and retrieval pipeline** NEW 70

Knowhere's progressive retrieval pipeline keeps a searchable baseline live while enhancements complete. It delivers immutable baseline and final snapshots with same-job completion milestones, durable recovery, and failed-enhancement retry, then atomically publishes the final result.

– Mechanism well described but no invocable command or flag given. v1.2.34 v1.2.32

THINNER COVERAGE BELOW

02 **Citation identity preservation across pipeline** NEW 40

Preserves physical source-page citation identity across the enhancement pipeline and retains explicit historical reads.

– Behavior described but no concrete usage detail provided. v1.2.34

03 **Page-memory TOC ownership** NEW 30

Adds page-memory table-of-contents (TOC) ownership for navigating large document collections.

– Named but not explained beyond a one-line description. v1.2.32

WAS THIS USEFUL? 👍 👎

rabitq-library

Feedback ⓘ v0.7.2 NOTES ↗ RANK

An official lightweight library for the RaBitQ algorithm and its applications in vector search.

> rabitq-library's v0.7.2 release focuses on polishing the Python API, improving input handling, object representations, and editor support rather than adding core algorithmic capabilities.

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

01 **Python API usability improvements** IMPROVED 54

Python index methods now accept path-like objects as inputs, with improved error messages shown on invalid arguments. Index objects such as `IvfIndex` and `HnswIndex` gained concise string representations, and the Python bindings now ship with editor type stubs (type hints) and accompanying API documentation.

— Named classes given but no code sample or exact signature shown v0.7.2

WAS THIS USEFUL? 👍 👎

Neo4j Agent Memory

Feedback ⓘ python-v0.8.0 CODE ↗ RANK

Agent Memory is a graph-native memory system for AI agents that stores conversations, builds knowledge graphs, and tracks reasoning traces with Neo4j.

> This window shipped two breaking changes: removal of deprecated feedback-tool keyword arguments and a bump of the minimum supported Python version to 3.11.

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

01 NAMS feedback tool keyword arguments removed

BREAKING

60

The `user_identifier` and `user_id` keyword arguments are removed from NAMS feedback tools in MCP, Pydantic AI, and Strands; callers passing these keywords must omit them.

– Names exact removed arguments and affected integrations python-v0.8.0

THINNER COVERAGE BELOW

02 Python 3.10 support dropped

BREAKING

50

Python 3.10 is no longer supported; `requires-python` is now `>=3.11` across the Python and NAMS AI provider packages.

– Names exact config key and version but no migration detail python-v0.8.0

nams-ai-provider-v0.4.0

└──▷ **BREAKING ON UPGRADE**

! The `user_identifier` and `user_id` keyword arguments are removed from NAMS feedback tools in MCP, Pydantic AI, and Strands; callers passing these keywords must omit them.

! Python 3.10 support is removed; `requires-python` is now `>=3.11`.

! Python 3.10 is no longer supported; `requires-python` is now `>= 3.11`.

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

Feedback ⓘ snapshot-20261028 NOTES ↗

RANK

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

> LangSmith's biggest addition this window is AI-powered natural-language search across traces and threads, alongside MCP server tunnel/OAuth improvements, GitHub App token support for Agent Auth, a new deployment repo-update API, and several breaking changes to deployment UI and Teams API access.

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

01 AI-powered search for traces and threads NEW 80

Tracing project pages now support AI-powered natural-language search across runs, traces, and threads. It is toggled per organization by an Organization Admin in `Organization Settings > Projects` via an Enable AI Search control that requires reviewing a data-processing and PHI consent step; thread search also supports an `exclude_metadata` field (e.g. to exclude automation-sourced threads) via `POST /threads/search`.

Exclude automation threads (e.g. from Open SWE sidebar) when searching threads so only user-initiated threads appear in results.

```
$ curl -X POST https://api.smith.langchain.com/threads/search \
  -H 'Authorization: Bearer <API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"exclude_metadata": [{"source": "automation"}]}'
```

Enable AI search for your organization so team members can query runs and traces in plain language from the Projects page.

> 📍 In LangSmith, go to Organization Settings › Projects, click Enable AI Search, review the data-processing and PHI consent step, and confirm as an Organization Admin.

— Names admin path, consent step and search field with runnable example.

snapshot-20261028

02 Update deployment source repo via API NEW 80

Deployments can now be pointed to a renamed or transferred GitHub repository without tearing down and recreating them, via `PATCH /v2/deployments/<id>` with a `source_config.repo_url` field.

Point an existing deployment to a renamed or transferred GitHub repository without tearing down and recreating it.

```
$ curl -X PATCH
https://api.smith.langchain.com/v2/deployments/<id> \
  -H 'Authorization: Bearer <API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"source_config": {"repo_url": "https://github.com/my-
org/new-repo-name"}}'
```

— Exact endpoint, method and field with runnable curl example. snapshot-20261028

03 **MCP server tunnel and OAuth token support** IMPROVED

68

Custom MCP servers behind a LangSmith tunnel now accept headers such as `x-api-key` and `Cookie` exactly as servers outside a tunnel, enabling OAuth-protected MCP servers to use a tunnel for token, registration, and protected-resource endpoints; authorization endpoints placed behind a tunnel are rejected with a `400` error at provider creation and update. The hosted LangSmith MCP server also now accepts access tokens issued for the LangSmith API, enabling a single OAuth grant to cover both the API and MCP.

— Names headers, error code and endpoints but no runnable command. snapshot-20261028

04 **GitHub App installation tokens via Agent Auth** NEW

60

Agent Auth now mints GitHub App installation tokens, allowing agents to call `connections.get` on a `github_app` connection and receive a renewable token.

— Names method and connection type but no full example. snapshot-20261028

05 **Deployment UI and settings changes** BREAKING

60

The separate Revisions tab in Deployment is removed, so bookmarks to tabs after Overview now open the tab to their right; custom headers for a deployment are no longer shown in the Settings dialog and now live exclusively in the Connect dialog; and deleting a deployment now requires typing `permanently delete` to confirm.

— Describes exact UI behavior changes but no code surface. snapshot-20261028

THINNER COVERAGE BELOW

06 **Teams API gated by can_use_teams setting** BREAKING 45

Teams APIs now return `403` unless the organization has the `can_use_teams` setting enabled.

— Names setting and status code but lacks further mechanism. snapshot-20261028

└──▷ **BREAKING ON UPGRADE**

! The separate Revisions tab in Deployment is removed; bookmarks to tabs after Overview now open the tab to their right.

! Teams APIs return `403` unless the organization has the `can_use_teams` setting enabled.

! Custom headers for a deployment are no longer shown in the Settings dialog; they now live exclusively in the Connect dialog.

! Deleting a deployment now requires typing `permanently delete` to confirm.

WAS THIS USEFUL? 👍 👎

Comet Opik

Feedback ⓘ 2.2.128 NOTES ↗ RANK

Opik traces, evaluates, and monitors LLM applications and agentic workflows, with datasets, an LLM-as-judge metric library, and production dashboards.

> Opik's latest updates focus on Playground prompt management and smarter provider-specific model settings, including automatic handling of unsupported sampling parameters for newer Gemini and Vertex AI models.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 **Auto-drop unsupported sampling params for Gemini 3+/Vertex AI**

IMPROVED 55

Automatically drops sampling settings and penalties that Gemini 3+ and Vertex AI models refuse, preventing rejected requests.

— Explains mechanism and named models but no config surface 2.2.124

02 **Playground variant and prompt management actions** NEW

50

The Playground gains a duplicate-variant action to clone any existing prompt variant, a blank Add variant action to start a new variant from scratch, and load and save actions for both Text and Chat prompt types.

— Names concrete UI actions but no mechanism detail beyond that 2.2.128

03 **Generic model settings panel for Anthropic and OpenAI** NEW

45

Adds a generic settings panel drawn from the model catalog for the Anthropic and OpenAI providers in the UI.

— Names providers and UI location but no field-level detail 2.2.124

WAS THIS USEFUL? 👍 👎

Langfuse

Feedback ⓘ v4.71.0 NOTES ↗ RANK

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

> Langfuse v4.71.0 adds self-hosted government licensing, richer AI gateway observability and analytics, and small UI refinements to Topics and gateway team notifications.

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

01 **Self-hosted government plan via license keys** NEW 57

Adds self-hosted government plan support via `lf_gov_` license keys for enterprise air-gapped or regulated deployments.

– Names exact license key prefix but lacks setup steps v4.71.0

02 **AI gateway observability facts and analytics tracking** NEW 52

Records model, key owner, byte count, and upstream error facts on AI gateway generations for richer observability, and adds product analytics tracking for AI gateway setup, credentials, invites, keys, and model access events.

– Lists concrete fields and tracked events but no mechanism detail v4.71.0

03 **Show unassigned toggle in Topics** NEW 48

Adds a Show unassigned toggle to the Topics actions menu to surface unclassified traces.

– Clear UI location but no further behavior explained v4.71.0

04 **Individual recipient pills for AI gateway teams** IMPROVED 40

Replaces the notify-all toggle in AI gateway team recipient selection with individual member pills.

– Names the UI change but omits rationale or migration impact v4.71.0

WAS THIS USEFUL? 👍 👎

overmind

Feedback ⓘ v0.1.19 NOTES ↗ RANK

The platform for continuously improving AI agents.

> Overmind v0.1.19 adds OTLP/JSON trace ingestion, surfaces Claude and OTel GenAI span content in its observability view, and makes OCR document processing more resilient to unreadable pages.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 **OTLP/JSON ingestion for OpenTelemetry traces** NEW 50

Overmind now accepts `OTLP/JSON` as an ingestion format for OpenTelemetry traces, expanding ingest beyond the previously supported wire formats.

— Names the format but no ingestion endpoint or config detail given. v0.1.19

02 **Claude and OTel GenAI span visibility** NEW 50

The observability view now surfaces Claude and OTel GenAI span content, making LLM-specific span data visible within the platform.

— Describes what's visible but not exact UI path or data fields. v0.1.19

03 **Resilient OCR batch processing for scanned documents** IMPROVED 50

Overmind reads photographed or scanned documents via OCR and continues processing a batch even when an individual page yields no text.

— Explains behavior change but no config or limits specified. v0.1.19

WAS THIS USEFUL? 👍 👎

Braintrust

Feedback ⓘ snapshot-20261028 NOTES ↗ RANK

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

> Braintrust's October 2026 release adds Cursor CLI/desktop tracing, new Loop integrations for PagerDuty and GitHub, dataset and human-review workflows in Loop/MCP, several dashboard additions including SQL charts and a built-in Coding agent insights dashboard, continuous Loop automations, a redesigned thread view, customizable Loop Patterns severity, and SDK updates for Go and TypeScript.

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

01 Dashboards NEW

90

Dashboards gain SQL charts built from a SQL query with a toggle between query text and a visual builder, also creatable in the SQL sandbox and saved to a dashboard (Pro and Enterprise plans; requires data plane v2.1.1+); Loop can now be opened from an individual dashboard or chart editor to investigate data and build or refine charts using natural language; and a new built-in Coding agent insights dashboard, found under `Dashboards > Coding agent insights`, shows usage by user, API key, model, tool, and branch with cost breakdowns, plus a daily Loop automation that surfaces recurring problems (requires data plane v2.13.0+).

– Names plan tiers, version, and the SQL sandbox entry point Clear UI entry point but no mechanism or version detail Exact navigation path, named breakdown dimensions, and version

snapshot-20261028

02 Cursor agent session tracing NEW

85

The `bt trace enable cursor` command (requires `bt` v0.26.0+) traces Cursor CLI and desktop app agent sessions, capturing turns, model calls, tool calls, and subagents.

Set up tracing for Cursor agent sessions so every turn, model call, and tool call is captured in Braintrust.

```
$ bt trace enable cursor
```

– Exact command and version requirement given, with a runnable example snapshot-20261028

03 Human review and dataset workflows in MCP/Loop NEW

70

The Braintrust MCP server and Loop now support sending project log traces, experiment traces, or dataset rows to human review and optionally assigning them to reviewers

(requires data plane v2.19.0+), and copying project log rows into datasets while preserving input, metadata, tags, and source log links (requires data plane v2.18.0+).

— Both workflows named with version gates but no exact commands `snapshot-20261028`

04 **Customizable severity and categories for Loop Patterns**

IMPROVED

70

Loop Patterns now supports a rubric letting teams define their own categories and severity criteria per project (requires data plane v2.19.0+), plus per-pattern severity and category assignments that can be corrected with an optional reason Loop preserves in later investigations (requires data plane v2.18.0+).

— Mechanism and versions named but no exact config path shown `snapshot-20261028`

05 **Loop automations enhancements** NEW

65

Loop automations gain a new PagerDuty notification destination supporting log, time-window, and environment alerts, configurable with an integration key and incident severity per destination (requires data plane v2.17.0+), and a continuous mode where Loop checks for new traces between runs and analyzes them in the current thread once enough traces arrive (requires data plane v2.20.0+).

— Names configuration fields and version but no exact setup steps Mechanism described but no exact configuration surface named

`snapshot-20261028`

THINNER COVERAGE BELOW

06 **Redesigned thread view with summaries** NEW

50

A new thread view (public preview) organizes a trace by conversation turn, showing the request, agent work, and final response per turn; it also adds thread summaries with Loop-generated headings per turn and a whole-conversation summary with click-to-jump navigation.

— Describes UI behaviour but no exact navigation or config given `snapshot-20261028`

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI/LLM Security

containarium

Feedback ⓘ v0.120.1 NOTES ↗ RANK

Open-source agent runtime — SSH-native isolation, eBPF egress policy, Kubernetes + LXC backends, GPU passthrough, MCP-native CLI

> Containarium's v0.120.x releases add a human-in-the-loop permission-request flow for agent code runs, signer key rotation with trust bundles and RFC 3161 timestamp anchoring, RWX shared volumes with CSI snapshot backup destinations, and backup-safety/peer-visibility controls — while removing the bridge-DNS subsystem entirely.

└──▷ **WHAT SHIPPED · 12 FEATURES** most completely described first ? what's the number?

01 Human-in-the-loop permission requests for agent code runs

NEW 95

Adds a `--approvals ask` flag to `code run`, with `code pending` and `code answer` subcommands, to pause a run on a permission request; pushes waiting permission requests to a phone via the `ntfy` integration; and requires a `code:answer` scope to answer permission requests.

Pause an agent code run when it hits a permission gate, then approve it from a second terminal — gives a human in the loop before the agent proceeds.

```
$ # Start a run that pauses on permission requests
containarium code run --approvals ask <run-id>

# In another terminal, inspect the pending request
containarium code pending <run-id>

# Approve it
containarium code answer <run-id>
```

— Full flag, subcommands and scope with a runnable example. v0.120.0

02 Signer rotation, trust bundles and timestamp anchoring NEW

90

Adds `signer rotate`, `signer revoke`, and `signer bundle` subcommands plus `verify --trust-bundle` to manage and validate signer trust bundles; adds an RFC 3161 timestamp-authority anchor sink for `verify-anchor`; and adds `signer revocation` and `expected_revision` fields on the server policy.

Rotate a compromised signer and publish a new trust bundle so relying parties can verify with the updated chain.

```
$ containarium signer rotate
containarium signer bundle
containarium verify --trust-bundle <bundle-file>
```

— Named subcommands, flags and fields with runnable example. v0.120.0

03 **MCP server integration for AI agent clients** NEW 75

Provides an MCP server (`mcp-server`) that can be wired into tools like Claude Code via `~/.claude.json` , configured with `CONTAINARIUM_SERVER_URL` and `CONTAINARIUM_JWT_TOKEN` env vars so an agent can manage containers directly.

Wire the platform MCP into Claude Code so an AI agent can manage containers directly from your laptop.

json

```
{
  "mcpServers": {
    "containarium": {
      "command": "/usr/local/bin/mcp-server",
      "env": {
        "CONTAINARIUM_SERVER_URL": "http://your-host:8080",
        "CONTAINARIUM_JWT_TOKEN": "<your-token>"
      }
    }
  }
}
```

— Exact config file and env vars shown, runnable. v0.120.0

04 **Shared volumes and snapshot backup destinations** NEW 74

Adds an RWX shared volume backend for read-write-execute multi-attach volumes, adds `define CSI snapshot backup destination` for snapshot-based backup targets, and refuses creates and snapshots past a pool headroom floor.

— Names volume backend, config action and enforcement rule. v0.120.0

05 **Bridge-DNS subsystem removed** BREAKING 71

Removes the bridge-DNS reconciler, record writer, status RPC, CLI subcommand, MCP tool, and alert; bridge-DNS flags are now no-ops.

— Names every removed component and migration note. v0.120.0

06 **SSH key replace-all RPC and cross-box lookup** NEW 62

Adds a `SetSSHKeys` replace-all RPC to atomically replace a box user's full set of authorized keys, and adds a cross-box lookup to find which box connected to a destination.

— Names RPC and new cross-box lookup capability. v0.120.0

THINNER COVERAGE BELOW

07 **Tenant status and access scoping** IMPROVED 54

Extends `status` to report whether a tenant runs the host default or a stored policy, and guards native network Tenant membership while scoping atomic Tenant access updates.

— Names status field and access-control behavior. v0.120.0

08 **SDLC skill set selection by id** IMPROVED 52

Adds `select SDLC skill sets by id` instead of the `sdlc_skills` boolean, and supports multi-source skill locks.

— Names old boolean and new selection, scope unclear. v0.120.0

09 **Backup verification and integrity controls** NEW 48

Adds backup verification controls: the ability to flag or refuse plaintext dumps backed by a backup status preflight check, a `VerifyProofPack` RPC so a client can verify a sealed pack, `GetTrackerStatus` reporting the credential kind as a typed enum, and a `backups:verify` scope for verifier tokens restricting which tokens can perform backup verification.

— Mechanism described but no command or flag named. Names RPC and enum but no further detail. Names the RPC but no further mechanism given. Names the scope but no further detail.

v0.120.1 v0.120.0

10 **Plain-HTTP peer state visibility** NEW 29

Makes plain-HTTP peer state visible for monitoring unencrypted peer connections.

— Short description, no mechanism or surface named. v0.120.1

11 **Structured host security posture reporting** NEW 19

Adds structured host security posture reporting.

— Bare description with no mechanism or fields named. v0.120.0

12 **Task node resource validation before scheduling** NEW 16

Validates local task node resource requirements before scheduling.

— Bare statement, no mechanism or limits given. v0.120.0

└──▷ **BREAKING ON UPGRADE**

! The bridge-DNS reconciler, record writer, status RPC, CLI subcommand, MCP tool, and alert are removed; bridge-DNS flags are now no-ops.

WAS THIS USEFUL? 👍 👎

NVIDIA OpenShell

Feedback ⓘ v0.1.6 NOTES ↗ RANK

OpenShell runs autonomous AI agents in policy-controlled sandboxes with isolated filesystem, network, and credential access.

> OpenShell v0.1.6 adds sandbox suspend/resume, an experimental shared Kubernetes supervisor, OCSF audit events for gateway security actions, and new Helm chart fields, alongside several breaking changes to gateway authentication, the sandbox-supervisor control connection direction, and the external isolation backend API.

└──▷ **WHAT SHIPPED · 9 FEATURES** most completely described first ? what's the number?

01 Experimental shared Kubernetes supervisor topology NEW

90

Setting `supervisor_topology = "gateway"` under `[openshell.drivers.kubernetes]` enables an experimental shared supervisor that consolidates many sandboxes into a single process, with replica count configurable via `supervisor.replicas` for failover capacity.

Enable the experimental shared Kubernetes supervisor to host many sandboxes in one process and configure two replicas for failover capacity.

toml

```toml
[openshell.drivers.kubernetes]
supervisor_topology = "gateway"

[supervisor]
replicas = 2
```

— Config keys and runnable example given; explicitly marked experimental v0.1.6

02 Sandbox suspend and resume commands NEW

77

New `openshell sandbox suspend` and `openshell sandbox resume` commands pause and restore sandboxes on supported compute drivers, preserving in-memory state so a long-running agent task can be paused mid-task and resumed later.

Pause a long-running agent sandbox mid-task and restore it later, preserving in-memory state.

```
$ openshell sandbox suspend my-agent-sandbox
# ... later ...
openshell sandbox resume my-agent-sandbox
```

— Exact commands and workflow given, internals of state preservation unclear v0.1.6

03 **Mandatory gateway authentication** BREAKING 75

Gateways now refuse to start unless OIDC, mTLS user or operator authentication, or gateway JWTs are configured; for trusted local development only, set `allow_unauthenticated_users = true` under `[openshell.gateway.auth]`.

— Exact config key and escape hatch given v0.1.6

04 **New Helm chart configuration fields** NEW 73

The Helm chart gains `server.extraEnv` for passing extra environment variables—including Kubernetes Secret references—to the gateway container, plus `pkiInitJob.podLabels` and `pkiInitJob.podAnnotations` for labeling and annotating certificate hook pods.

Pass AWS credentials as a Kubernetes Secret reference to the gateway container via Helm without hardcoding values in values.yaml.

```
$ helm upgrade --install openshell deploy/helm/openshell \
  --namespace openshell \
  --set 'server.extraEnv[0].name=AWS_ACCESS_KEY_ID' \
  --set 'server.extraEnv[0].valueFrom.secretKeyRef.name=aws-
creds' \
  --set 'server.extraEnv[0].valueFrom.secretKeyRef.key=access-
key-id'
```

— Field names and one helm command given; podLabels/podAnnotations usage not shown v0.1.6

05 **ResumeBoundaryRequest field renamed** BREAKING 68

For external isolation backends, `ResumeBoundaryRequest` field 3 is renamed from `allow_cold_start_fallback` to `restart_if_restore_fails`.

— Exact field names given, narrow scope v0.1.6

06 **Reversed sandbox-supervisor control connection direction** 67

BREAKING

New sandboxes on Kubernetes, Docker, Podman, and VM drivers now open the control connection from sandbox to supervisor, the opposite direction from before; older supervisor and sandbox runtime binaries reject new sandboxes, so both must be upgraded together.

— Clear before/after and upgrade requirement, no command given v0.1.6

THINNER COVERAGE BELOW

07 **Prerelease builds on snap beta channel** NEW 58

Prerelease builds now publish to the snap's `latest/beta` channel; running `OPENSHELL_VERSION=pre install.sh` installs from there.

— Exact install command given, minimal further detail v0.1.6

08 **OCSF audit events for gateway security actions** NEW 57

Gateways now emit OCSF audit events for authorization denials, accepted credentials, SSH session tokens, and sandbox token checks, and raise a High-severity finding on suspected sandbox credential misuse.

— Event types named but no endpoint or query interface given v0.1.6

09 **Enforced middleware binding policy** NEW 50

Operators can enforce middleware binding requirements by setting `binding_policy` to `allowlist` or `exact`.

— Config values named but no file path or enforcement mechanism shown v0.1.6

└──▷ **BREAKING ON UPGRADE**

! Gateways now refuse to start unless OIDC, mTLS user or operator authentication, or gateway JWTs are configured; for trusted local development only, set `allow_unauthenticated_users = true` under `[openshell.gateway.auth]`.

! New sandboxes on Kubernetes, Docker, Podman, and VM open the control connection from sandbox to supervisor in the opposite direction from before; older supervisor and sandbox runtime binaries reject new sandboxes, so both must be upgraded together.

! External isolation backends using `ResumeBoundaryRequest`: field 3 is renamed from `allow_cold_start_fallback` to `restart_if_restore_fails`.

WAS THIS USEFUL? 👍 👎

cplt

Feedback ⓘ 2026.10.27-154720-48d556d NOTES ↗ RANK

cplt is a kernel-level sandbox for AI coding agents, running Copilot CLI, Claude Code, OpenCode, Gemini CLI, and others with git and gh guards.

> cplt extended its macOS sandboxing to cover brokered MCP server child processes and added an opt-in to let opencode skip project config in unapproved repositories.

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

01 Seatbelt sandbox covers brokered MCP child processes IMPROVED

57

The macOS Seatbelt/SBPL sandbox, enforced via `sandbox-exec`, now extends to brokered MCP server child processes spawned by the MCP broker.

— Names mechanism and scope but no usage path 2026.10.27-113426-7287cb3

02 Opencode opt-in to skip unapproved-repo config NEW

40

Adds opt-in behaviour for `opencode` to skip project config in unapproved repositories.

— States the behaviour but not how to opt in 2026.10.27-154720-48d556d

WAS THIS USEFUL? 👍 👎

agent-safe-pipeline

Feedback ⓘ v0.5.0 NOTES ↗ RANK

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

> AgentSafe v0.5.0 adds subject-bound scoping across sessions and MCP tool access, new gateway-level policy controls including a facts-only privacy mode, and governance primitives for composite plans, write preconditions, and contradiction handling.

└──▷ **WHAT SHIPPED · 7 FEATURES** most completely described first ? what's the number?

01 **Gateway policy enforcement controls** NEW 57

The gateway gains a `facts-only` privacy mode that keeps the full request body out of the authority log, per-HTTP-route policy selection so different published policies can be enforced on different gateway routes, and an opt-in enforcement mode for hosted tenants moving beyond shadow-only evaluation.

– Named `facts-only` mode but no config key or flag shown v0.5.0

02 **Subject-bound scoping for sessions and MCP tools** NEW 57

Adds subject-bound scoping across the platform: sessions that scope an agent's reads to one context to prevent cross-context data leakage, and declarative tool profiles read by `McpGuard` that apply subject-bound scoping to MCP tool access.

– No profile schema or session API detail given v0.5.0

03 **Precondition enforcement on state version** NEW 55

Adds precondition enforcement that binds a target's state version and refuses a write on version mismatch.

– Mechanism clear but no API or flag named v0.5.0

04 **Action space manifest format** NEW 50

Introduces the `agent-safe.actions/1` manifest format for writing out the action space.

– Format named but no schema or fields documented v0.5.0

05 **Local authority re-derivation for testing** NEW 40

Adds local authority re-derivation at the claim for testing, including modelling held allowances in the local authority.

– Testing mechanism named but no command shown v0.5.0

06 **Observed-contradiction outcome handling** NEW 40

Adds observed-contradiction outcome handling that holds its scope and supports resume flows.

— Behavior described but no trigger or API named v0.5.0

07 **Governed plans for composite decisions** NEW 25

Adds governed plans for composite decisions, implementing ADR 0008.

— Bare reference to ADR, no mechanism shown v0.5.0

WAS THIS USEFUL? 👍 👎

beelzebub

Feedback ⓘ v3.10.0 NOTES ↗

Beelzebub is a low-code deception runtime that runs AI-powered honeypots simulating systems for attackers.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v3.10.0 https://github.com/beelzebub-labs/beelzebub.git
# already have the repo? check out this version:
$ git checkout v3.10.0
```

NEW DETECTION WordPress 6.9.4 service lure

NEW DETECTION Ollama service lure

WAS THIS USEFUL? 👍 👎

opensandbox

Feedback ⓘ release-1.2.0 NOTES ↗ RANK

Secure, Fast, and Extensible Sandbox runtime for AI agents.

> OpenSandbox 1.2.0 ships auto-pause and wake-on-access for fast sandboxes with a matching wake-retry contract, a Credential Vault API for sandbox secrets, OCI registry publishing for snapshots, and a new diagnostics context endpoint. It also adds pinned MITM CA certificates per pool, a structured renewal policy replacing a deprecated extend key, port-scoped egress rules, unbounded SDK file transfers, and an activity-suppressing header, across the server and all five SDKs.

└──▷ **WHAT SHIPPED · 9 FEATURES** most completely described first ? what's the number?

01 Auto-pause and wake-on-access for fast sandboxes NEW

95

Adds `lifecycle.idlePolicy` create field (`idleTimeoutSeconds` 30–86400, `onIdle: pause` , `wakeOnAccess` , optional `activeProbe`) enabling sandboxes to automatically pause after inactivity and wake transparently on the next request; server-side config moves under `[automation.idle_pause]` with `redis.dsn` validated at load when the sweeper is enabled, and all five SDKs add client-side fail-fast validation for the typed field. A companion wake-retry contract has SDKs resend after a `503` carrying the `OpenSandbox-Wake: expired` header with Retry-After , up to a bounded budget (default 5 attempts / 60s), raising a typed wake-timeout exception on exhaustion; stream-start wake-retry landed in JavaScript, Python, and Go.

Create a fast sandbox that automatically pauses after 5 minutes of inactivity and wakes transparently on the next request.

```python
import asyncio
from opensandbox import Sandbox
from opensandbox.models import IdlePolicy, RenewalPolicy

async def main() -> None:
    sandbox = await Sandbox.create(
        "python:3.12",
        lifecycle={
            "idlePolicy": {
                "idleTimeoutSeconds": 300,
                "onIdle": "pause",
                "wakeOnAccess": True,
            }
        },
    )
    print(f"Sandbox {sandbox.id} created with auto-pause
enabled")
    await sandbox.destroy()

asyncio.run(main())
```

Enable the server-side idle sweeper so paused fast sandboxes are reclaimed automatically, with Redis tracking activity.

```toml
[automation.idle_pause]
enabled = true

[redis]
dsn = "redis://localhost:6379/0"

[kubernetes]
template_mitm_ca_secret = "pool-mitm-ca"
```

— Named config keys, field ranges, and a runnable example `release-1.2.0`

02 **Credential Vault API for sandbox secrets** NEW 90

Adds `POST|GET|PATCH|DELETE /v1/sandboxes/{id}/credential-vault` API (OSEP-0022) where credential values never appear on public reads, `PATCH` uses `expectedRevision` as a cross-replica backstop, and non-fsb backends proxy the sidecar's `/credential-vault` surface; all five SDKs add a Credential Vault client contract, and template-backed sandboxes route vault calls through the server on the same template-origin signal used for network policy.

— Endpoints and revision mechanism named, no usage example release-1.2.0

03 **Pinned MITM CA certificates per Kubernetes pool** NEW

83

Adds `[kubernetes] template_mitm_ca_secret` server config (default off) that pins the MITM CA per pool by emitting `spec.egressMitmCertificate` on every SandboxTemplate, so an egress restart no longer rotates the trust anchor under live sandboxes.

Enable the server-side idle sweeper so paused fast sandboxes are reclaimed automatically, with Redis tracking activity.

toml

```toml
[automation.idle_pause]
enabled = true

[redis]
dsn = "redis://localhost:6379/0"

[kubernetes]
template_mitm_ca_secret = "pool-mitm-ca"
```

— Config key and mechanism named with a config example release-1.2.0

04 **Sandbox diagnostics context endpoint** NEW

80

Adds `GET /v1/sandboxes/{sandboxId}/diagnostics/context` endpoint returning sandbox ID, lifecycle state, backend type, and backend-defined identifiers for Docker, Kubernetes, and Fast Sandbox, without requiring an execd connection — useful for triage when a pod has not yet been allocated.

Inspect a sandbox's lifecycle state and backend identifiers without an active execd connection — useful for triage when a pod has not yet been allocated.

```
$ curl -s -H 'Authorization: Bearer <your-api-key>' \
  https://<server>/v1/sandboxes/<sandboxId>/diagnostics/context |
jq .
```

— Exact endpoint with a runnable curl example release-1.2.0

05 **Structured renewal policy replaces legacy extend key** BREAKING

78

Adds `lifecycle.renewalPolicy.extendSeconds` (300–86400) as the structured replacement for the opaque `extensions['access.renew.extend.seconds']` key,

which is now deprecated and rejected by fast sandbox with a 400 instead of silently accepting a no-op; the legacy key is supported through 1.x and removed in 2.0.

— Exact migration path and deprecation timeline named `release-1.2.0`

THINNER COVERAGE BELOW

06 **Activity-suppressing sentinel header** NEW 55

Adds `OpenSandbox-Skip-Activity` sentinel header that suppresses one activity observation and is stripped on both gateway and proxy hops.

— Exact header name given, no usage example `release-1.2.0`

07 **OCI registry publishing for Fast Sandbox snapshots** NEW 40

Supports publishing Fast Sandbox snapshots and template builds to — and booting from — OCI registries via the OverlayBD OCI lifecycle.

— No config keys or commands given for this flow `release-1.2.0`

08 **Unbounded SDK file upload and download transfers** IMPROVED 35

Removes whole-call timeouts from SDK file uploads and downloads in all five languages, eliminating mid-flight failures on large or slow transfers.

— No flag or API name given, prose only `release-1.2.0`

09 **Port-scoped egress rules in classic sandbox profile** NEW 25

Adds port-scoped egress rules end to end in the classic sandbox profile.

— Thin description with no named config or API `release-1.2.0`

└──▷ **BREAKING ON UPGRADE**

! The `extensions['access.renew.extend.seconds']` key is deprecated and fast sandbox now rejects it with 400 instead of silently accepting a no-op; migrate to `lifecycle.renewalPolicy.extendSeconds` (supported throughout 1.x; the legacy key is removed in 2.0).

WAS THIS USEFUL? 👍 👎

ToolHive

Feedback ⓘ v0.58.0 NOTES ↗ RANK

ToolHive runs MCP servers in isolated containers and enforces identity and access policies for local and Kubernetes deployments.

> ToolHive v0.58.0 strengthens its embedded Authorization Server with RFC 9068-compliant token typing and tighter controls over stale refresh sessions.

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

01 **AT+JWT token typing for embedded AS tokens** NEW 55

Adds an opt-in `at+jwt` `typ` header for embedded Authorization Server access tokens, enabling RFC 9068-compliant JWT typing for downstream consumers.

— Names the exact header value and RFC but no config flag given v0.58.0

02 **Refresh grant lifecycle enforcement** IMPROVED 35

The embedded Authorization Server now refuses refresh grants whose upstream credential has expired or been revoked, preventing stale delegated sessions from persisting.

— Describes behavior but no named surface or trigger mechanism v0.58.0

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Anthropic

Feedback ⓘ changelog · Oct 28, 2026 CHANGELOG ↗ RANK

The Claude Developer Platform is the API for building applications on Anthropic's Claude models that generate, analyze, and automate work.

> Anthropic expanded administrative and safety controls across the Claude Developer Platform — new Claude Tag, organization webhook, and user-profile enforcement endpoints, organization-wide web fetch controls, new Claude Managed Agents session capabilities, and a new API version with major SDK releases — alongside new model releases including Claude Opus 5.5 and the retirement of Opus 4.1 and Sonnet 4.

└──▷ **WHAT SHIPPED · 12 FEATURES** most completely described first ? what's the number?

01 **User profiles and enforcement status API** NEW 90

Adds `/v1/user_profiles` endpoints so platforms and gateways can tell Anthropic which end customer each API request is for, enabling safety enforcement to target individual profiles rather than the whole organization. Adds `enforcement_status` and `enforcement_actions` fields on user profiles plus an `enforcement_status[]` filter on the list-profiles endpoint to surface whether Anthropic has restricted or banned an end customer.

Check whether any of your end-customer profiles are currently under an Anthropic restriction or ban before processing their requests.

```
$ curl 'https://api.anthropic.com/v1/user_profiles?
enforcement_status[]=restricted&enforcement_status[]=banned' \
  -H 'x-api-key: <admin-api-key>' \
  -H 'anthropic-version: 2026-10-21'
```

— Named endpoint, fields and filter with a runnable curl example

changelog-20261028-30823831

02 **Claude Tag endpoints for Slack governance** NEW 80

Adds `/v1/organizations/claude_tag/` endpoints (beta, `experimental-claude-tag-2026-10-21` header) to read/write Claude Tag settings on Slack channels and workspaces, show effective settings and their sources, and attach Access bundles. Authenticated with Admin API keys scoped to `read:claude_tag` or `write:claude_tag`.

— Names endpoint, header and scopes but no HTTP methods shown changelog-20261028-30823831

03 **API version 2026-10-21 and major SDK releases** BREAKING

75

Released API version `2026-10-21`, which new major SDK versions default to: Python SDK 2.0.0, TypeScript SDK 1.0.0, C# SDK 13.0.0, Go SDK 2.0.0, Java SDK 3.0.0, PHP SDK 1.0.0, Ruby SDK 2.0.0, and CLI 2.0.0. These releases carry breaking changes independent of the API version bump: all SDKs and the CLI are now licensed under Apache 2.0 instead of MIT, and page cursors issued without the new pagination header become invalid once the header is adopted, requiring callers to restart from the first page.

– Exact versions and license change named, but no migration command

changelog-20261028-30823831

04 **Organization-wide web fetch controls** NEW

75

Adds controls for web fetch on the Capabilities page of the Claude Console: admins can disable web fetch entirely for their organization, after which requests using the web fetch tool fail with a 400 error, and can set domain filtering via allowed or blocked domains that apply across all Messages API requests and Claude Managed Agents sessions.

– Exact UI path and failure behavior named, no API flag given changelog-20261028-30823831

05 **Organization webhooks for user restrictions** NEW

70

Adds `/v1/organizations/webhooks` endpoints (beta, `webhooks-2026-10-07` header) to register webhook endpoints for two event types: when Anthropic restricts or bans a user profile, and when a restriction or ban ends.

– Endpoint and header named, event types listed, no payload shown

changelog-20261028-30823831

06 **New Claude Managed Agents session capabilities** NEW

65

Adds a per-session spend budget capped at public list rates that pauses the session when reached, advisor support letting the session's primary thread consult a more-capable model mid-turn for strategic guidance, the ability to restrict which sites an agent's web tool can fetch (configurable on the tool's entry), and credential injection that securely passes secrets into the agent's sandbox via environment variables for CLIs, SDKs, and services that authenticate through environment variables.

– Mechanism explained for each capability but no exact config keys

changelog-20261028-30823831

07 **Thinking block compatibility restrictions across models**

65

BREAKING

Thinking blocks produced by Claude Sonnet 5.5 are only usable within the account that produced them or a linked account; sending them from another account causes the API to silently drop the block. On Claude Opus 5.5, thinking cannot be disabled. Thinking blocks from Claude Fable 5.1 and Claude Mythos 5.1 are dropped by the API when replayed to an earlier model.

— Names specific models and behavior but gives no workaround changelog-20261028-30823831

08 **API key expiration notices and reporting** NEW 60

Adds API key expiration support: Anthropic emails the creator before expiration for keys with a lifetime of at least 7 days, and the Admin API reports each key's expiration date.

— Concrete threshold and reporting field named, no exact API field

changelog-20261028-30823831

THINNER COVERAGE BELOW

09 **Claude Models on Google Cloud** NEW 50

Launches Claude Models on Google Cloud (Anthropic managed) — the Claude API operated by Anthropic, accessed through a Google Cloud project, with Google Cloud IAM for access control and Google Cloud Marketplace for billing.

— Describes the offering but no setup steps given changelog-20261028-30823831

10 **Observability tracing for Messages API and agents** NEW 45

Adds observability tracing (beta) so Messages API requests and Claude Managed Agents sessions emit spans in the Claude SDK for Python, TypeScript, C#, Go, Java, PHP, and Ruby.

— States what emits spans but no setup or config detail changelog-20261028-30823831

11 **Admin API member management for Claude Enterprise** NEW 45

Expands the Admin API with member management endpoints for Claude Enterprise: list members, look them up by email, change roles, remove members, send and withdraw invites, manage groups and membership, and read custom roles.

— Lists capabilities but no endpoint paths or methods given changelog-20261028-30823831

12 **Opus 4.1 and Sonnet 4 model retirements** BREAKING 30

Claude Opus 4.1 model has been retired. Claude Sonnet 4 model has been retired.

— Bare retirement notice with no migration guidance changelog-20261028-30823831

└──▷ BREAKING ON UPGRADE

- ! The new Python SDK 2.0.0, TypeScript SDK 1.0.0, C# SDK 13.0.0, Go SDK 2.0.0, Java SDK 3.0.0, PHP SDK 1.0.0, Ruby SDK 2.0.0, and CLI 2.0.0 include breaking changes independent of the API version bump.
- ! All SDKs and CLI are now licensed under Apache 2.0 instead of MIT.
- ! Thinking blocks produced by Claude Sonnet 5.5 are only usable within the account that produced them or a linked account; sending them from another account causes the API to silently drop the block.
- ! On Claude Opus 5.5, thinking cannot be disabled.
- ! Thinking blocks from Claude Fable 5.1 and Claude Mythos 5.1 are dropped by the API when replayed to an earlier model.
- ! Claude Opus 4.1 model has been retired.
- ! Claude Sonnet 4 model has been retired.
- ! Page cursors issued without the new pagination header are not valid when the header is adopted; callers must restart from the first page.

WAS THIS USEFUL? 👍 👎

Mistral AI

Feedback ⓘ snapshot-20261028 NOTES ↗ RANK

Learn how to deploy and use Mistral AI's Large Language Models with our comprehensive documentation, guides, and tutorials.

> Mistral AI added domain filtering for its web search tool, letting developers restrict or allow specific sources queried by the model.

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

01 **Web search domain filtering** NEW 33

Mistral AI's web search capability now supports domain filtering, allowing developers to restrict or allow specific domains when the model performs web searches.

— Describes capability but no config key or API field named `snapshot-20261028`

WAS THIS USEFUL? 👍 👎

xAI Grok API

Feedback ⓘ snapshot-20261028 NOTES ↗ RANK

Grok API provides xAI's Grok models for building applications with chat, image and video generation, voice, tool calling, and structured outputs.

> Grok API added custom content moderation controls, letting developers set per-category handling rules for NSFW, violence, and profanity.

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

01 **Custom moderation guidelines array** NEW 65

A new `guidelines` array can be added to Chat Completions and Responses API requests to set per-category content handling rules for the `nsfw`, `violence`, and `profanity` categories.

– Names exact field and categories but no values/schema shown snapshot-20261028

WAS THIS USEFUL? 👍 👎

Google Gemini API

Feedback ⓘ October 27, 2026 NOTES ↗

RANK

Gemini API provides Gemini models for generating text, images, and other multimodal content in applications.

> Gemini API introduced a beta command-line client for scripting and CI use against its Interactions API.

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

01 **gemini-api CLI in beta** NEW 55

The `gemini-api` CLI launches in beta as a stateless, zero-dependency command-line client for the Interactions API, letting users generate text and media, analyze files, and run Gemini managed agents from scripts and CI pipelines.

– Names the CLI and its capabilities but no flags or commands October 27, 2026

WAS THIS USEFUL? 👍 👎

Upstage Solar Pro 3 Feedback ⓘ

retirement scheduled · Oct 27, 2026 MODEL ↗

Solar Pro 3 is a chat model from Upstage.

OpenRouter lists a retirement date of 30 Oct 2026 for upstage/solar-pro-3, first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

◆ Open-Weight Models

ConvAI Innovations Laya Feedback ⓘ v0.4.16 NOTES ↗

RANK

Non-autoregressive System 1 decision engine. Typed choice, score and yes/no decisions over any text in a single forward pass, in 100+ languages, with a router that picks the right checkpoint per request.

> Laya v0.4.16 brings batch prediction and confidence-gating parity to its .NET and Rust SDKs.

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

01 **Batch prediction APIs in .NET and Rust SDKs** NEW 57

Adds `predict_batch` and router batch APIs to both the .NET SDK and the Rust SDK, letting callers submit multiple decision requests in a single call instead of one at a time.

— Names exact methods but no usage example or limits given v0.4.16

02 **Confidence-gating in .NET and Rust SDKs** NEW 57

Adds a `min_confidence` abstention gate and `jev_confidence` value to both the .NET SDK and the Rust SDK, allowing callers to suppress or flag low-confidence decisions.

— Names exact config fields but no behavior detail or example v0.4.16

WAS THIS USEFUL? 👍 👎

Liquid AI LFM3-1.7B

Feedback ⓘ released · Oct 28, 2026

MODEL ↗ DOCS ↗

LFM3-1.7B from Liquid AI was released on 23 Oct 2026. Open weights on Hugging Face.

Liquid AI adds Omni Models (text, image, video, audio in/out) and new LFM3/d1 models with context windows up to 256K tokens.

Introduces Omni Models category supporting text, image, video, and audio input with text or speech output in a single model.

Adds `LFM3-1.7B` with a 256K-token context window, `LFM3-VL-2B` with a 128K-token context window, `LFM3-Audio-2B` (coming soon), and `LFM3-Omni-2B`, an omni model supporting text, image, video, and audio.

Adds `d1-2B` with a 128K-token context window (coming soon), `d1-omni-2B` with a 128K-token context window (coming soon), `d1-omni-600M` with a 16K-token context window and runtime support for Transformers, llama.cpp, and mlx-vlm, and `d1.1`.

+1 more

WAS THIS USEFUL? 👍 👎

Liquid AI LFM3-Omni-2B

Feedback ⓘ

released · Oct 28, 2026 MODEL ↗

LFM3-Omni-2B from Liquid AI was released on 21 Oct 2026. Open weights on Hugging Face.

WAS THIS USEFUL? 👍 👎

Liquid AI LFM3-VL-2B

Feedback ⓘ released · Oct 28, 2026

MODEL ↗

LFM3-VL-2B from Liquid AI was released on 19 Oct 2026. Open weights on Hugging Face.

WAS THIS USEFUL? 👍 👎

PrismML Ternary Bonsai 2 9B

Feedback ⓘ

released · Oct 28, 2026 MODEL ↗

Ternary Bonsai 2 9B from PrismML was released on 16 Oct 2026. 262K-token context; $0.08 in / $0.13 out per million tokens (OpenRouter).

WAS THIS USEFUL? 👍 👎

Liquid AI d1-2B

Feedback ⓘ released · Oct 28, 2026 MODEL ↗

d1-2B from Liquid AI was released on 24 Oct 2026. Open weights on Hugging Face.

WAS THIS USEFUL? 👍 👎

Liquid AI d1-Omni-2B

Feedback ⓘ released · Oct 28, 2026

MODEL ↗

d1-Omni-2B from Liquid AI was released on 24 Oct 2026. Open weights on Hugging Face.

WAS THIS USEFUL? 👍 👎

DeepSeek V4 Flash 0731 Feedback ⓘ

price increase · Oct 27, 2026 MODEL ↗

DeepSeek V4 Flash 0731 is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4-flash-0731): $0.0444 → $0.119 in, $0.1333 → $0.238 out per million tokens (+168%), first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

DeepSeek V4.1 Flash Feedback ⓘ

price increase · Oct 27, 2026 MODEL ↗

DeepSeek V4.1 Flash is a multimodal model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4.1-flash): $0.02 → $0.14 in, $1.00 → $0.42 out per million tokens (+600%), first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

Google Gemma 3 4B Feedback ⓘ price increase · Oct 27, 2026

MODEL ↗

Gemma 3 4B is a multimodal model from Google.

Price increase on OpenRouter (google/gemma-3-4b-it): $0.04 → $0.05 in, $0.05 → $0.10 out per million tokens (+100%), first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

Moonshot AI Kimi K2.7 Code

Feedback ⓘ

price cut · Oct 27, 2026 MODEL ↗

Kimi K2.7 Code is a coding model from Moonshot AI.

Price cut on OpenRouter (moonshotai/kimi-k2.7-code): $0.6712 → $0.56 in, $3.35 → $3.10 out per million tokens (-17%), first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

Xiaomi MiMo-V2.5

Feedback ⓘ price increase · Oct 27, 2026

MODEL ↗

MiMo-V2.5 is a multimodal model from Xiaomi.

Price increase on OpenRouter (xiaomi/mimo-v2.5): $0.119 → $0.40 in, $0.238 → $2.00 out per million tokens (+740%), first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

Xiaomi MiMo-V2.5-Pro

Feedback ⓘ

price increase · Oct 27, 2026 MODEL ↗

MiMo-V2.5-Pro is a chat model from Xiaomi.

Price increase on OpenRouter (xiaomi/mimo-v2.5-pro): $0.3045 → $0.48 in, $0.609 → $1.80 out per million tokens (+196%), first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Models

Feedback ⓘ

qwen-audio-3.1-realtime-flash · NOTES ↗ · RANK

Qwen is Alibaba's language and multimodal model family, including coding, vision, image, speech, and reasoning models.

> Alibaba added qwen-audio-3.1-realtime-flash, a new real-time duplex speech model to the Qwen family.

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

01 **Real-time duplex speech model qwen-audio-3.1-realtime-flash**

NEW 65

`qwen-audio-3.1-realtime-flash` is a new real-time duplex speech model supporting text and audio input and output, function calling, web search, and voice cloning, with a 262,144-token context window.

— Names model and capabilities but no usage example or API details.

qwen-audio-3.1-realtime-flash

WAS THIS USEFUL? 👍 👎

Alibaba Qwen3.5-27B

Feedback ⓘ

price increase · Oct 27, 2026 · MODEL ↗

Qwen3.5-27B is a multimodal model from Alibaba.

Price increase on OpenRouter (qwen/qwen3.5-27b): $0.27 → $0.26 in, $2.16 → $2.60 out per million tokens (+20%), first seen 27 Oct 2026.

WAS THIS USEFUL? 👍 👎

Black Forest Labs FLUX

Feedback ⓘ

FLUX 3 Image: Custom Output Sizes and Better Edits · NOTES ↗ · RANK

FLUX is Black Forest Labs' family of visual generation and editing models.

> FLUX 3 Image adds custom output sizing and resolution-matching controls to its editing API, alongside accuracy improvements for single-reference-image edits.

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

01 Custom output sizing and resolution controls in FLUX 3 Image

NEW

95

Adds `width` and `height` parameters to `POST /v1/flux-3-image`, accepting any size from 256px per side up to 16.8MP (4K) at aspect ratios up to 8:1 (4MP limit when reference images are included); `width` and `height` override `resolution` and `aspect_ratio`. Adds `resolution: "match_input"` for edits, outputting at the same resolution as the first reference image up to 4MP, with aspect ratio honoured at a standard size if set. Also removes cropping on single-reference-image edits when `aspect_ratio` is set to `auto`, using the full input image instead.

— Names exact endpoint, params and size limits; fully actionable.

FLUX 3 Image: Custom Output Sizes and B…

THINNER COVERAGE BELOW

02 Improved accuracy for single-reference-image edits

IMPROVED

45

Improves edit accuracy for single-reference-image edits, covering whole-image changes such as weather, relighting, color grading, restoration, text edits, and camera moves — no request changes required.

— Describes scope of improvement but no mechanism or metric given.

FLUX 3 Image: Custom Output Sizes and B…

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · October 28, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue