NOW AVAILABLE The 512GB Mac Studio is on sale now, shipping February to March. See every configuration.

THE AI TOOLCHAIN

NO. 085

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED NOVEMBER 6, 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 387 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 94 of 95 tools

▾ // THE BRIEF

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

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

vim ? help

DEPLOY

Run a model that doesn't fit on one Mac by pairing two of them

mlx-serve adds tensor parallelism across two Thunderbolt 5-linked Macs and brings GGUF weights onto its faster MLX engine; Darkbloom adds GPU memory tuning, external-disk support and automatic version parity between paired Macs. Local inference stops being bounded by a single machine's unified memory, and the pair stops drifting out of sync on updates.

mlx-serve · Darkbloom

BUILD

Hand a cloud agent a brief and get back a running app with a database behind it

Cursor's cloud agents scaffold, provision Postgres, run an automated security review and publish, with the Agent SDK now GA and a PR-replay command so webhook integrations can be tested against a recorded event instead of a fresh PR. flow-next closes the other end: `flow --auto --until=merge` drives GitLab merge requests through review to merge without a human relaying state.

Cursor · flow-next

GOVERN

Bound what an agent can read and what it can spend before it runs

OKF Agent Memory makes memory-layer access declarative per bundle, so an agent physically cannot read context it wasn't scoped for, and a lockfile-plus-checksum pass stops two concurrent writers from corrupting shared memory. Hermes tracks spend on paid MCP tools, turning per-call tool costs into something visible before the invoice.

OKF Agent Memory · Hermes

Does **Opus 5** read better? 👍 👎

BUILD

◆ AI Coding Agents

Anysphere Cursor

Feedback ⓘ updated · Nov 6, 2026 NOTES ↗

RANK

Cursor is a built to make you extraordinarily productive, agents turn ideas into code. Accelerate development by handing off tasks to Cursor.

> Cursor's cloud agents can now scaffold, provision and publish live web apps end-to-end — complete with automated security review, Postgres databases, and team sharing — while the Agent SDK graduates to general availability with a new PR-replay command for webhook testing.

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

01 agent-sdk github replay command for webhook fixtures NEW

90

The `agent-sdk github replay <pr-url>` command turns a pull request into signed webhook payloads on push; adding `--dry-run --fixtures-dir fixtures/` saves the payloads as eval fixtures for testing agent behaviour.

Save signed webhook payloads from a PR as local eval fixtures to reproduce agent behaviour in tests.

```
$ agent-sdk github replay https://github.com/your-org/your-repo/pull/42 --dry-run --fixtures-dir fixtures/
```

— Exact runnable command with flags and purpose snapshot-20261106

02 Live app deployment from cloud agent sessions NEW

85

Cloud agent sessions can build and deploy live web apps directly: Cursor scaffolds the app, spins up a Postgres database if needed, configures authentication behind Cursor SSO, provisions hosting from a natural-language prompt, deploys from new or existing repos across Cursor Origin and GitHub, and publishes to a custom domain by picking a unique URL, returning a live link once checks pass.

Kick off a cloud agent session and ask it to build and deploy a live internal tool for your team.

📍 1. Open Cursor and start a cloud agent session. 2. Type: 'Build a live web app that shows our sales pipeline — connect it to our data warehouse and deploy it.' 3. When prompted, provide the required secrets (e.g. warehouse credentials). 4.

When the agent finishes, choose a unique domain and click Publish to get a shareable link.

— Mechanism and named surfaces detailed, but flow is UI-driven not a command

snapshot-20261106

03 **Automated security review for app publishing** NEW 70

Before any app publish, Cursor runs automatic deterministic and agentic security checks that catch issues like secrets written into app files or exposed credentials; nothing goes live until checks pass and the user explicitly publishes. Team admins can view every app and secret on the team, add custom rules to the security review, approve publishes that the review blocked, and unpublish any app.

— Clear mechanism and admin controls named, no exact command or path snapshot-20261106

THINNER COVERAGE BELOW

04 **Sharing published apps with teams or publicly** NEW 45

Published apps can be shared with specific people, groups, or the entire team behind Cursor sign-in; individual plan users can also share apps publicly.

— States sharing scope but no concrete UI path or config snapshot-20261106

05 **API key access to deployed apps** NEW 35

A Cursor API key can be provisioned to allow a script or Grok Bot to use a deployed app.

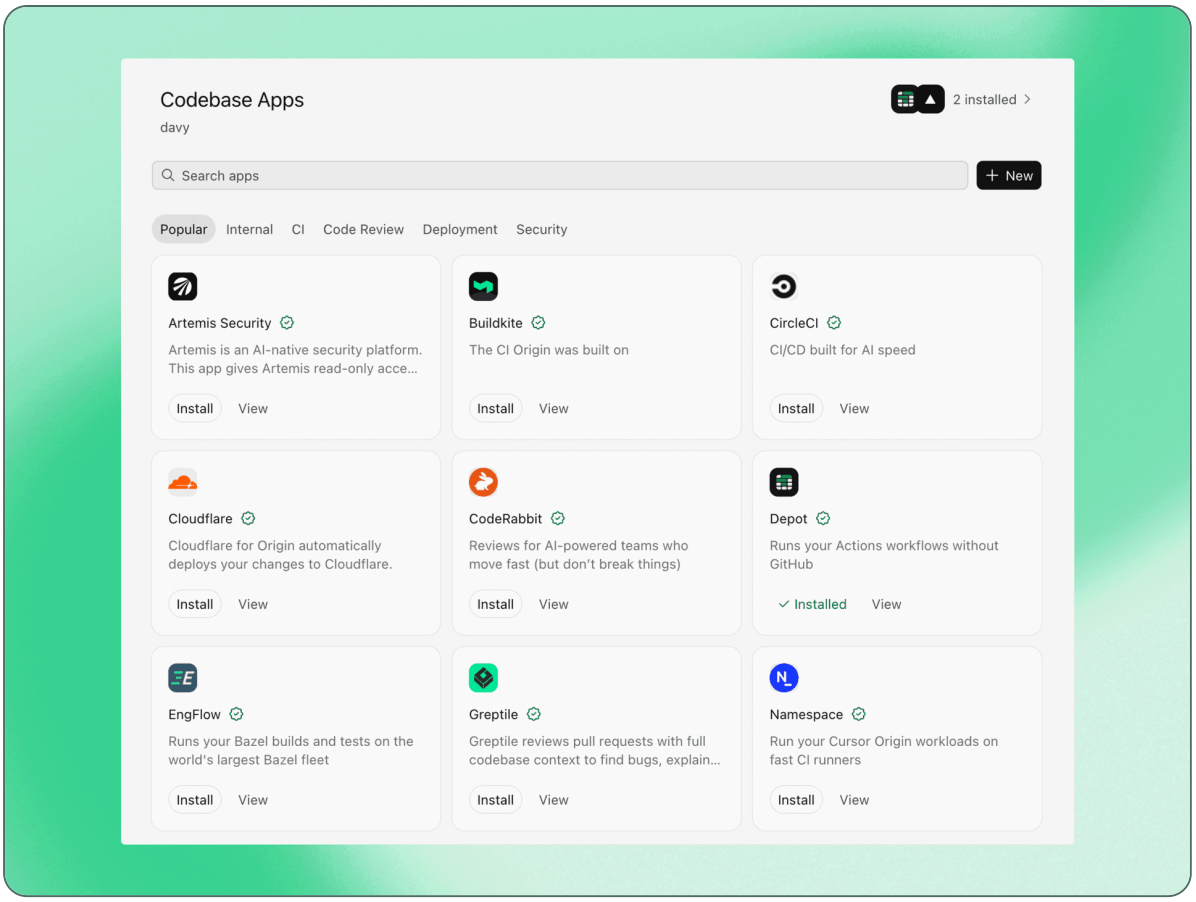
— Names the key and Grok Bot but no setup detail snapshot-20261106

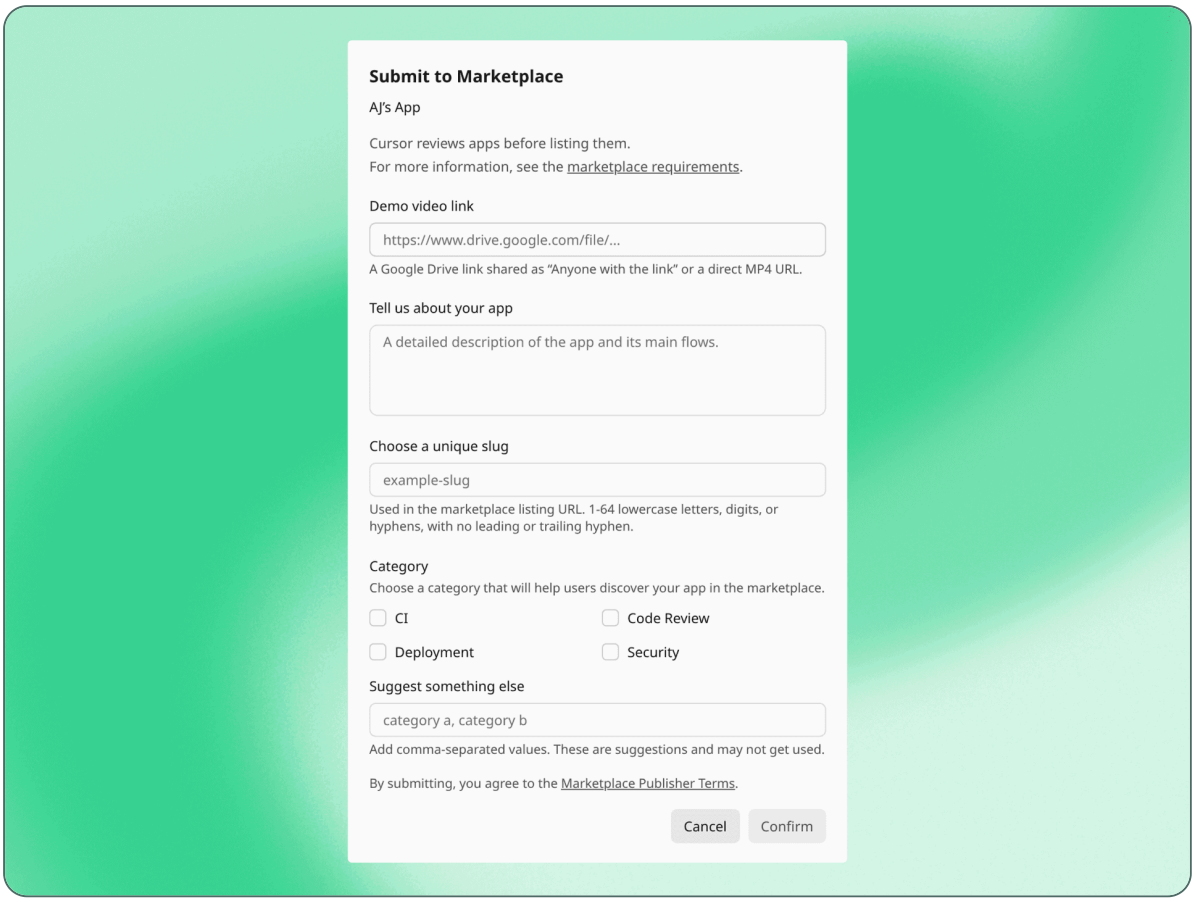
06 **Agent SDK reaches general availability** IMPROVED 20

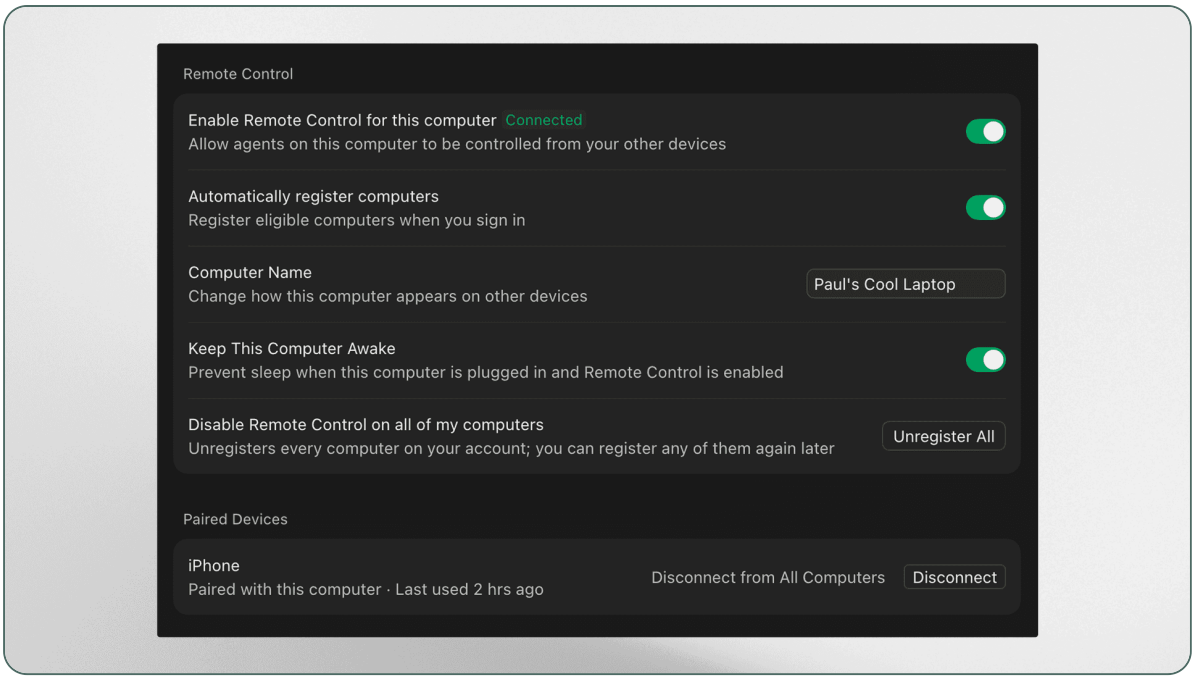
The Agent SDK is now generally available.

— Bare status announcement with no further detail snapshot-20261106

└──▷ ALSO FROM THESE RELEASES


Codebase Apps
davy
2 installed
Search apps
New
Popular
Internal
CI
Code Review
Deployment
Security
Artemis Security
Artemis is an AI-native security platform. This app gives Artemis read-only acce...
Install
View
Buildkite
The CI Origin was built on
Install
View
CircleCI
CI/CD built for AI speed
Install
View
Cloudflare
Cloudflare for Origin automatically deploys your changes to Cloudflare.
Install
View
CodeRabbit
Reviews for AI-powered teams who move fast (but don't break things)
Install
View
Depot
Runs your Actions workflows without GitHub
Installed
View
EngFlow
Runs your Bazel builds and tests on the world's largest Bazel fleet
Install
View
Greptile
Greptile reviews pull requests with full codebase context to find bugs, explain...
Install
View
Namespace
Run your Cursor Origin workloads on fast CI runners
Install
View


Submit to Marketplace
AJ's App
Cursor reviews apps before listing them.
For more information, see the marketplace requirements.
Demo video link
https://www.drive.google.com/file/...
A Google Drive link shared as "Anyone with the link" or a direct MP4 URL.
Tell us about your app
A detailed description of the app and its main flows.
Choose a unique slug
example-slug
Used in the marketplace listing URL. 1-64 lowercase letters, digits, or hyphens, with no leading or trailing hyphen.
Category
Choose a category that will help users discover your app in the marketplace.
CI
Code Review
Deployment
Security
Suggest something else
category a, category b
Add comma-separated values. These are suggestions and may not get used.
By submitting, you agree to the Marketplace Publisher Terms.
Cancel
Confirm



WAS THIS USEFUL?

yetone magpie

Feedback ⓘ v0.1.3140 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's biggest moves this window turn it into a multi-instance MCP hub: a remote magpie's MCP router can now be registered as a Library server and reached from headless or gateway-mode deployments, gated behind a new per-provider switch. Alongside that, terminal launching and launch-line copying extend to Windows and Linux, custom providers gain HTTP/1.1 pinning for broken HTTP/2 connections, and OpenCode gets explicit prompt-cache-key forwarding.

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

01 Remote magpie MCP router exposed as Library server NEW

93

A remote magpie's MCP router is now automatically registered as a Library server named for its provider, available at that instance's `gateway.HubMCPAll` endpoint with the provider's key, so its tools appear in the hub as `<name>__<tool>` entries and can be toggled per agent. Remote-router Library servers are marked `Bridge: true` and excluded from tool listings when a request arrives from another magpie's router (detected via `RouterHeader` / `FromRouter`), preventing circular tool-chain loops between two magpie instances that each list the other as a remote provider.

— Names endpoint, header, flags and tool-naming scheme precisely v0.1.3127

02 HTTP/1.1 pinning for custom providers NEW

88

A custom provider can be pinned to HTTP/1.1 by setting the `HTTP1` field on its provider record, causing requests to that provider's host to use a dedicated HTTP/1.1 transport while all other hosts continue to use HTTP/2. magpie also automatically demotes a provider's host to HTTP/1.1 when a TLS record error occurs on an HTTP/2 connection, without requiring manual configuration.

— Exact config field and automatic fallback mechanism given v0.1.3133

03 **OpenCode prompt cache key forwarding** NEW 82

magpie's OpenCode provider config now sets `setCacheKey: true` in the provider `options` block, so OpenCode 1 (via camelCase `promptCacheKey`) and OpenCode 2.1.4+ (via `compatibility.supportsPromptCacheKey`) forward the session's `prompt_cache_key` to upstream on every request; a user-set value, including `false`, is preserved across syncs.

— Exact config keys and version-specific behaviour named v0.1.3129

04 **Per-provider switch gating for remote magpie MCP servers** 80

BREAKING

Remote magpie MCP servers now require an explicit per-provider switch to be turned on before they are added to the Library, replacing the previous 'offered once on add' behaviour. The switch, offer, and server state recorded under a provider's old ID are automatically moved to the new ID when a provider is renamed. Providers added before this release keep their switch shown as on by default; providers added after start with the switch off and must be turned on before their MCP server enters the Library, and removing a server from the Library turns its switch off so it won't be re-offered until re-enabled.

— Clear before/after and migration behaviour for existing providers v0.1.3135

05 **Open in terminal and copy launch line on Windows/Linux** NEW 78

The open in terminal capability, previously macOS-only, now works on Windows (Windows Terminal, PowerShell 7 via `pwsh.exe`, or Windows PowerShell via `powershell.exe`) and Linux (GNOME Terminal, Konsole, and others). The Sessions page also gains an action to copy any CLI agent's launch line, producing PowerShell-encoded output on Windows and a POSIX shell line on Linux.

— Names exact terminals, binaries and the Sessions page action v0.1.3137

06 **MCP router access for headless and gateway-mode deployments** NEW 68

In gateway mode, remote clients can reach the MCP router using a gateway key, while connections without a valid key are refused, matching the access-control behaviour of the `/v1` endpoint. This lets headless (no-agent) and gateway-mode deployments connect remote apps to all Library servers at one address.

— Names gateway key and matching endpoint but lacks exact address format v0.1.3115

THINNER COVERAGE BELOW

07 **Text selection on routing group cards** IMPROVED 50

Routing group cards now allow text — patterns, model names, and group IDs — to be selected and copied directly on the card or in its editor, without triggering the editor-open click handler.

— Clear UI change but no config or API surface v0.1.3134

08 Faint lower-ranked models in usage speed chart IMPROVED

42

The usage speed chart now plots every model that has timed data at each point, not just top-ranked models, drawing lower-ranked models in a faint color so they remain visible without crowding the legend, across both 7-day and 30-day views.

— Describes visual change but no concrete interaction point v0.1.3140

└──▷ **BREAKING ON UPGRADE**

! Remote magpie providers added before this release that already had a Library server will have their switch shown as on; providers added after this release start with the switch off and must be turned on before their MCP server enters the Library. Removing a server from the Library now turns the switch off, so the server will not be re-offered until the switch is turned back on.

WAS THIS USEFUL? 👍 👎

ai-rulez

Feedback ⓘ v5.11.0 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 v5.11.0 adds leakage scanning for imported text with configurable areas-file filters, and extends the exposure command with a Codex harness budget and an opt-in MCP tool-schema cost estimate.

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

01 **Codex harness budget and MCP tool-schema estimate in exposure command**

NEW 55

The exposure command now adds a Codex harness budget and an opt-in MCP tool-schema cost estimate.

— Names the exposure command but lacks flag syntax v5.11.0

02 **Imported-text leakage scanning with areas-file filters** NEW

45

ai-rulez now scans imported text for leakage and supports `exclude` and `accept` filters configured via an areas file.

— Names areas file and filter types but no exact config key v5.11.0

WAS THIS USEFUL? 👍 👎

zot

Feedback ⓘ v0.4.66 NOTES ↗ RANK

zot manages coding agent workflows from the command line.

> zot improved extension readability with session aliases and added native support for Gemini custom providers.

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

01 **Native Gemini custom provider support** NEW 25

zot now supports native Gemini custom providers, expanding the built-in provider list.

— Names provider but no config details given v0.4.66

WAS THIS USEFUL? 👍 👎

keploy

Feedback ⓘ v3.6.221 NOTES ↗ RANK

Open-source platform for creating safe, isolated production sandboxes for API, integration, and E2E testing.

> Keploy improved its mock replay engine with better handling of unmatched parser calls and recordings that persist across repeated test requests.

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

01 Parser call-miss handling and diagnostics IMPROVED

35

Keploy's agent parsers now hold unmatched calls at the end of a test set rather than discarding them, reducing missed-match noise during replay. The agent also now names parsers whose set-end filing has not returned, making incomplete parser states identifiable in diagnostics.

— Describes mechanism but no named flags or config to act on v3.6.220 v3.6.221

02 Recording reuse for repeated requests during replay IMPROVED

35

Keploy now keeps a test's recording for repeated requests, serving the stored response on the first match so repeats are covered without re-recording. Recordings marked not required are now served without being consumed during mock replay, preserving them for reuse.

— Clear behavior change but no concrete commands or settings given v3.6.220 v3.6.221

WAS THIS USEFUL? 👍 👎

GitHub Copilot CLI

Feedback ⓘ v1.0.106 NOTES ↗ RANK

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

> GitHub Copilot CLI's v1.0.106 release adds a `/code` command for jumping into VS Code, a `COPILOT_SESSION_CWD` environment variable for LSP/MCP server launches, opt-in session management with `@mentions`, a new GPT-6.1 Terra model, and tightens security defaults around device-code login and plugin auto-apply rules.

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

01 **`COPILOT_SESSION_CWD` environment variable** NEW 75

A new `COPILOT_SESSION_CWD` environment variable lets you set a fixed working directory for launching and expanding LSP and local MCP server configs, so relative paths resolve from that root rather than wherever `copilot` was invoked.

Set a working directory for LSP and MCP server launch so relative paths resolve correctly from a fixed root rather than wherever copilot was invoked.

```
$ export COPILOT_SESSION_CWD="/home/user/myproject"
copilot
```

— Exact env var and runnable export example given v1.0.106

02 **Expanded `copilot help` documentation** IMPROVED 75

`copilot help config` now lists every setting in the settings catalog, including `askUser`, `effortLevel`, `planModel`, `toolSearch` and `worktreeBaseRef`; `copilot help permissions` documents glob patterns in write(path) and read(path) rules, lists the `read` and `memory` kinds, and explains that an unscoped `write` rule approves shell output redirection but not the command itself; `copilot help billing` documents the `/statusline` `quota` and `ai-used` items and their defaults; `copilot help limits` shows a status example at the real minimum AI credit limit and notes the final-call warning appears when 10% or less of the limit remains; and `copilot help monitoring` now says an `http://` OTLP endpoint disables export only when an OTLP CA certificate or client certificate/key variable is set.

— Every documented flag and behavior named verbatim, runnable via help commands v1.0.106

03 **`/code` command opens session in VS Code** NEW 60

The new `/code` command opens the current Copilot CLI session directly in VS Code to continue editing with full IDE tooling.

Open your current Copilot CLI session directly in VS Code to continue editing with full IDE tooling.

```
$ /code
```

— Named command with runnable example, minimal mechanism detail `v1.0.106`

---

04 **Session delete command safeguards fixed** IMPROVED 60

`/session delete --yes=false` and `/session delete --yes=<value>` no longer delete the current session, and `/session delete --local-only <id>` now previews that session instead of acting on it directly.

— Exact flags named, behavior change clear, no example shown `v1.0.106`

THINNER COVERAGE BELOW

05 **HTTPS required for device-code login** BREAKING 55

`copilot login --device-code` now refuses a cleartext `http://` host where it previously accepted one; an `https://` host must be used instead.

— Named flag and exact remediation, no broader context `v1.0.106`

---

06 **Plugin rules with alwaysApply: false require opt-in** BREAKING 50

Plugin rules with `alwaysApply: false` no longer apply automatically every session; they now stay out of the prompt until enabled in `/instructions`, reversing the previous default.

— Behavior reversal explained, enabling path named but not exact command `v1.0.106`

---

07 **Fusion tier and sandbox extensions config settings** NEW 50

`copilot config` gains a Fusion tier setting to control tier selection, and a Sandbox extensions setting to run extension processes inside the sandbox.

— Named settings but only one-line description of each `v1.0.106`

---

08 **Opt-in session management and @mentions** NEW 40

Adds opt-in session management tools and `@mentions` for local sessions that have opted in.

— Thin description, no command or UI path given `v1.0.106`

---

09 **Unified `/settings` dashboard with global search** IMPROVED

40

`/settings` now opens the unified dashboard for all users and adds global search across settings and local resources.

— Named command and UI path, little mechanism detail `v1.0.106`

---

10 **GPT-6.1 Terra model added** NEW

35

GPT-6.1 Terra is now selectable under New models when available to your account.

— Bare name with minimal context on access or capabilities `v1.0.106`

└──▷ **BREAKING ON UPGRADE**

! `copilot login --device-code` now refuses a cleartext `http://` host where it previously accepted one; switch to an `https://` host.

! Plugin rules with `alwaysApply: false` no longer apply automatically every session — they are now inactive until enabled in `/instructions`, reversing the previous default.

WAS THIS USEFUL? 👍 👎

OpenAI Codex CLI

Feedback ⓘ rust-v0.167.0 NOTES ↗ CODE ↗

RANK

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

> OpenAI Codex CLI's latest releases add an agent command center with cross-backend session switching, a Google Cloud Vertex AI onboarding wizard, expanded Model providers settings in the TUI, terminal image previews, and tighter exec-policy and MCP controls.

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

03 **Agent command center and backend session switching** NEW

75

Open the agent command center with `codex agents --no-daemon` to monitor multiple concurrent agent tasks; the command center and session picker now mark unread results, and users can switch between connected backends from the terminal session picker while keeping drafts across backend switches.

Monitor multiple concurrent agent tasks and spot unread results without leaving the terminal.

```
$ codex agents --no-daemon
```

— Names runnable command and UI behavior clearly rust-v0.167.0

05 **Google Cloud Vertex AI onboarding wizard** NEW 70

Enable the Google Cloud Vertex AI setup wizard with `gcp_setup_wizard` during TUI onboarding, supporting Application Default Credentials or a credential file, accessible under Sign in another way.

— Names wizard, auth modes and navigation path rust-v0.167.0

THINNER COVERAGE BELOW

09 **Image and visual previews in terminal** NEW 55

Preview local image attachments in the TUI composer on supported terminals and open viewed images via clickable transcript paths; image previews are also scoped to individual panes in tmux, and pet displays now render in fullscreen tmux panes with Kitty passthrough.

— Describes behavior but no command to invoke rust-v0.168.0-alpha.6 rust-v0.167.0

12 **System proxy opt-in for macOS and Windows** NEW 45

Opt into macOS or Windows system proxy settings from the `/experimental` settings screen.

— Thin one-line description with a clear location `rust-v0.167.0`

WAS THIS USEFUL? 👍 👎

swttch

Feedback ⓘ v0.35.3 NOTES ↗ RANK

Claude Code GUI plugin for JetBrains IDEs. (ex - Claude Code with GUI)

> swttch's only release this window added desktop-style cards to the chat view for sessions and MCP tool calls.

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

01 **Desktop-style cards in chat view** NEW 38

swttch now renders desktop-style session cards showing plan and wrap-up boxes with links in the chat view, and also displays MCP tool calls as desktop-style cards in the same chat view.

— Describes UI additions but no exact navigation or config detail `v0.35.3`

WAS THIS USEFUL? 👍 👎

All Hands AI OpenHands

Feedback ⓘ v1.36.0 NOTES ↗

RANK

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

> OpenHands v1.36.0 adds service-account identity for Canvas automations, broadens repository support to Bitbucket Data Center and Azure DevOps, and reworks the automation template setup and management experience.

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

01 Self-hosted GitHub template setup and matching IMPROVED

70

A setup guide now adds a Connect Git step with a tour on the MCP page to guide self-hosted GitHub template configuration. Self-hosted GitHub templates now guide users through `MCP` or token setup, templates open without displaying the missing-integration notice, templates matching integrations the user has already configured are suggested and promoted, the setup assistant now receives the instance's configured event sources when building automations, and a template's selected repository is now sent only to services that clone from its provider.

— Lists several concrete behavior changes with a UI starting point v1.36.0

THINNER COVERAGE BELOW

02 Bitbucket Data Center and Azure DevOps in repository picker

NEW

55

Automation template setup now lists Bitbucket Data Center and Azure DevOps repositories in the repository picker. Data Center templates now ship with extensions `0.39.0` and use the repository picker with extensions `0.40.0`.

— Names providers and exact extension versions v1.36.0

03 Service-account identity for Canvas automations NEW

45

Automations now run as a service account in Canvas, separating automation identity from individual user credentials.

— Explains the change but gives no UI path or command v1.36.0

04 Unified automation create, view, edit workflow IMPROVED

35

A unified create, view, and edit workflow now replaces the separate automation screens.

— Describes the change only in general terms `v1.36.0`

WAS THIS USEFUL? 👍 👎

claude-mem

Feedback ⓘ v15.10.0 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's latest release focuses on multi-provider flexibility and reliability, adding per-source observer routing, long-lived OAuth token support to prevent capture interruptions, and improved diagnostics, alongside smaller onboarding, mode-chaining, and cross-device sync improvements.

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

01 Long-lived OAuth token for capture continuity NEW

88

Supply `CLAUDE_CODE_OAUTH_TOKEN` in `~/.claude-mem/.env` to keep capture running when the stored Claude Code login expires, with paused sessions resuming once the credential is verified.

Keep capture running past Claude Code login expiry by supplying a long-lived OAuth token.

env

```
CLAUDE_CODE_OAUTH_TOKEN=<your-oauth-token>
```

— Exact env var, config file path, and example provided. v15.10.0

02 Per-source observer provider routing NEW

75

The new `CLAUDE_MEM_PROVIDER_BY_SOURCE` setting lets each session source route to its own observer provider — for example, Codex sessions to Codex and Claude Code sessions to Claude.

— Named config key with concrete example, no file path given. v15.10.0

03 Doctor command reports provider availability IMPROVED

65

`npx claude-mem doctor` now reports an unavailable observer provider and its worker dependencies instead of reporting all is well.

— Runnable command named, behavior change briefly described. v15.10.0

THINNER COVERAGE BELOW

04 Chainable mode overrides NEW

45

Mode overrides can now be combined so a language mode and chill stack together, for example `code--chill--ko`.

— Example syntax given but mechanism unexplained. v15.10.0

05 **Onboarding discovers existing work first** IMPROVED 30

Onboarding now leads with discovering existing work, rules, and focus from transcripts before setup proceeds.

— Vague flow description, no concrete steps or surfaces. v15.10.0

06 **Cross-device My Work sync in viewer** NEW 30

Viewer sync now shares My Work state between devices and recovers remote changes on main.

— No mechanism or UI path described. v15.10.0

WAS THIS USEFUL? 👍 👎

coffee-cli

Feedback ⓘ v3.9.2 NOTES ↗ RANK

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

> Coffee CLI added in-app update support for macOS and Linux, including automatic installer handling and a mirror-racing mechanism to speed up downloads from mainland China.

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

01 **In-app updates for macOS and Linux** NEW 70

A new update button downloads the installer to the Downloads folder, verifies its integrity, then opens the DMG on macOS or the folder on Linux. On Linux it automatically selects the DEB or RPM package to match the system.

— Describes concrete update flow but no CLI command or flag named v3.9.2

THINNER COVERAGE BELOW

02 **Mirror racing for China network downloads** NEW 50

Coffee CLI now reads version numbers and release notes from mainland China mirrors and races multiple download sources simultaneously, using whichever one finishes first and passes verification.

— Explains mechanism but no configurable surface named v3.9.2

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Code

Feedback ⓘ desktop-v0.26.8 NOTES ↗

RANK

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

> Qwen Code's TypeScript SDK gains multi-modal input, session resume/export, an experimental LSP, and a Zed extension, while Qwen subscriptions now work across CLI, VS Code, and Desktop; Desktop and Web Shell also pick up RPM packaging, remote-access toggling, and expanded channel/session monitoring, alongside a breaking settings-key rename from `disable*` to `enable*`.

└──▷ **WHAT SHIPPED · 19 FEATURES** most completely described first ? what's the number?

01 Settings and configuration additions NEW

85

Remote access can now be enabled or disabled directly from settings without restarting the daemon or editing a config file, a new `config` operator model catalog refresh switch lets operators toggle whether the model catalog refreshes, and a new `settings.env` field lets you configure environment variables such as API keys and proxy settings directly in settings so every session picks them up without shell exports.

Set API keys and proxy settings once in your config so every session picks them up without shell exports.

json

```
{
  "env": {
    "OPENAI_API_KEY": "sk-...",
    "HTTPS_PROXY": "http://proxy.corp.example:8080"
  }
}
```

— Clear UI path but no exact setting name given Named switch but mechanism and effect unclear Exact config key plus a runnable example config

desktop-v0.26.8 sdk-typescript-v0.1.28

02 Session resume, continue, and export in TypeScript SDK NEW

78

Sessions can be resumed by passing a session ID via the `resume` and `continue` options in the TypeScript SDK, and session history can be exported to markdown or HTML with the new `export` command.

— Names SDK options and export formats, lacks a worked example sdk-typescript-v0.1.28

03 Web Shell channel and session monitoring additions IMPROVED

65

Web shell adds a conversation activity view for active sessions, a dismissible Workflow chip and menu summary in the composer, a Channel event timeline for monitoring channel activity, an Add to Split View option on session menus, exposed GitHub channel configuration, new host mount entries, an aligned managed-session empty-state layout, direct links from channel events to their GitHub PR and Issue targets, and links from channel event rows to their corresponding sessions.

— Many named UI surfaces but only navigation, no commands desktop-v0.26.8 v0.26.8

---

04 **`extra_body` field for OpenAI-compatible providers** NEW

65

OpenAI-compatible providers now support an `extra_body` field to pass arbitrary request fields through to the provider API.

— Named field and use case, no example shown sdk-typescript-v0.1.28

---

05 **Settings keys renamed from `disable*` to `enable*`** BREAKING

65

Settings keys previously named with a `disable*` prefix are renamed to a positive `enable*` naming convention; existing configuration using the old names will break.

— Clear migration note but no list of specific renamed keys sdk-typescript-v0.1.28

---

06 **Token usage panel enhancements in Web Shell** IMPROVED

60

Web shell's token usage panel now shows request success rate, distinct colors per model, and plots per-request speed and failures.

— Concrete metrics named but only a UI panel view desktop-v0.26.8

---

07 **Mod prompt submission in headless and Ink runs** NEW

60

Mods can now submit their own prompts via `prompt.submit` hooks during headless runs and inside the Ink terminal interface, enabling fully automated mod-driven pipelines in both headless and interactive sessions.

— Names the hook and both run modes it applies to v0.26.8 desktop-v0.26.8

---

08 **Conversations working-directory root config in serve daemon** NEW

60

The serve daemon now exposes a configurable Conversations working-directory root, controlling where multi-workspace sessions are anchored.

— Explains purpose but no config key name given v0.26.8

THINNER COVERAGE BELOW

09 **ACP shared loop for additional turn types** IMPROVED

55

ACP now runs re-entered, channel, Goal-proposal, cron, and background-notification turns on the shared loop.

— Names turn types but no user-facing action described desktop-v0.26.8

10 **Multi-modal input across content generators** NEW

55

The SDK now supports multi-modal input — image, PDF, and audio — across all content generators.

— Named input types but no usage path given sdk-typescript-v0.1.28

11 **`FORK_MODE` for ProcessTransport** NEW

55

`ProcessTransport` now supports a `FORK_MODE` for Electron IPC integration.

— Named flag and transport but no usage example sdk-typescript-v0.1.28

12 **Role-aware action templates for GitHub bound sessions** NEW

50

Role-aware action templates for GitHub bound-session delivery let channels apply different templates depending on the acting role.

— Describes behavior but no template names or config path v0.26.8

13 **Concurrent batch CLI runner** NEW

45

A new concurrent runner lets batch CLI tasks run in parallel.

— Describes capability but no flag or command named sdk-typescript-v0.1.28

14 **Zed extension for Qwen Code agent server** NEW

45

A new Zed extension integrates the Qwen Code agent server with the Zed editor.

— Names the integration but no setup steps sdk-typescript-v0.1.28

15 **`contextWindowSize` configuration option** NEW

45

A new `contextWindowSize` configuration option controls the model's context window size.

— Named config key, no default or range given sdk-typescript-v0.1.28

16 **Qwen subscription support across CLI, VS Code, Desktop** NEW

45

Qwen Code now supports Qwen subscriptions across the CLI, VS Code extension, and Desktop app.

— Names three surfaces but no setup or pricing detail v0.26.8

17 **Concurrent workspace rewinds in managed-agent** IMPROVED 45

Managed-agent now allows concurrent workspace rewinds, removing the previous one-at-a-time constraint.

— States before/after but no mechanism or limit v0.26.8

18 **RPM installer packaging for Desktop** NEW 40

Qwen Code Desktop now builds and publishes RPM installer packages for Linux RPM-based distributions.

— Names packaging format only, no install steps given desktop-v0.26.8

19 **Experimental LSP support** NEW 30

Experimental LSP support adds code intelligence to Qwen Code.

— Bare description with no detail on scope or setup sdk-typescript-v0.1.28

└──▷ **BREAKING ON UPGRADE**

! Settings keys previously named with a `disable*` prefix are renamed to a positive `enable*` naming convention — existing config using the old names will break.

WAS THIS USEFUL? 👍 👎

termide

Feedback ⓘ 0.49.0 NOTES ↗ RANK

All-in-one terminal workspace for desktops and servers: editor with LSP, file manager with SFTP/FTP, terminal, git, database viewer and a coding agent in one zero-config static Rust binary.

> Termide 0.49.0 adds a built-in CSV/TSV table viewer, gives the coding agent persistent named terminal sessions plus a new tool to manage them, turns on automatic prompt caching for Anthropic and OpenRouter, and renames the agent's `bash` tool to `shell` as a breaking change.

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

01 **CSV/TSV table viewer in file manager** NEW 90

CSV and TSV files can be opened as a sortable grid via `F3`, `Open...`, a bookmark or a followed link; `Ctrl+E` switches between the table and the source editor. Delimiter (`,`, `;`, `Tab`, `|`) and header detection are auto-guessed and adjustable from the status bar or panel menu; `s` sorts a column, `y` copies a cell, and rows are never fully loaded into memory so files of any size open immediately.

— Names exact keys, delimiters and memory behaviour 0.49.0

02 **Persistent named agent terminal sessions** NEW 90

Passing `session` to the agent's `shell` tool runs commands in a named terminal that persists its directory, variables and background jobs between calls; passing `background` starts a long-running process (dev server, watcher, slow build) and returns immediately. The number of agent terminal sessions per session is capped with `[ai] max_terminal_sessions` in the global config (default: 10).

Limit agent terminal sessions to 3 per session in a shared or restricted environment.

toml

```toml
[ai]
max_terminal_sessions = 3
```

— Named tool params and config key with default and example 0.49.0

03 **Agent `bash` tool renamed to `shell`** BREAKING 90

The agent's `bash` tool is renamed to `shell`: `[ai.permissions.bash]` in the global config and project rules, `tools` entries in `ai/hooks.toml`, global agents' `tools:`, and session log call names all become `shell`; hook inputs now name the tool `"shell"` and external agents see it as `mcp__termide__shell`. An unedited `ai/tools/bash.md` is replaced by `ai/tools/shell.md`, an edited one is moved

there with the new default placed alongside it as `.new`, and project-level `bash` entries are not rewritten automatically — they are reported and continue to apply only through this release.

— Full before/after migration detail with every renamed surface `0.49.0`

---

04 **terminals agent tool for managing agent terminals** NEW

70

The new `terminals` agent tool lists, reads output from, types into, interrupts or closes agent-owned terminals, and can wait for a string, a shell exit code, or output to settle before returning — enabling the agent to drive REPLs, debuggers and prompts.

— Mechanism and capabilities named but no runnable example `0.49.0`

---

05 **app_attribution toggle for usage tracking** NEW

65

The `app_attribution` setting on a connection's page in AI settings controls whether termide identifies itself to usage-tracking endpoints (OpenRouter today); it is on by default and the toggle appears only for endpoints that support app attribution.

Turn off app attribution for a connection so termide does not identify itself in OpenRouter's per-app usage breakdown.

toml

```toml
# In AI settings, open the connection's page and set:
app_attribution = false
```

— Exact setting name, location and example provided `0.49.0`

THINNER COVERAGE BELOW

---

06 **Automatic prompt caching for Anthropic and OpenRouter**

IMPROVED

55

Anthropic and OpenRouter connections now request automatic prompt caching so each turn reads prior conversation from cache at a fraction of the input cost; cached tokens appear in the status bar under `↻`.

— Explains mechanism and indicator but no configuration step `0.49.0`

└──▷ ALSO FROM THESE RELEASES

.

└──▷ BREAKING ON UPGRADE

! The agent's `bash` tool is renamed to `shell`: `[ai.permissions.bash]` in the global config and in project rules, `tools` entries in `ai/hooks.toml` and global agents' `tools:`, and session log call names all become `shell`. Hook inputs now name the tool `"shell"`, and external agents see it as `mcp__termide__shell`. An unedited `ai/tools/bash.md` is replaced by `ai/tools/shell.md`; an edited one is moved there with the new default placed

alongside it as `.new`. Project-level `bash` entries are not rewritten automatically — they are reported and continue to apply only through this release.

WAS THIS USEFUL? 👍 👎

Token Optimizer

Feedback ⓘ v5.13.64 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's single release this window focused on reducing context waste with targeted output reads, richer checkpoint state, and more accurate usage metrics.

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

01 **Line/match-targeted expand reads** NEW 69

The `expand` command can now retrieve only specific lines or matches from an archived tool result using targeted line or match filters, avoiding the need to read the full output.

— Names exact command and filter mechanism v5.13.64

THINNER COVERAGE BELOW

02 **Higher threshold for long-session size tips** IMPROVED 50

Size tips for long sessions now appear only above 300K tokens, keeping sessions below that threshold free of quality-score penalties.

— Names exact 300K token threshold v5.13.64

03 **Richer checkpoint fields for compaction recovery** IMPROVED 45

Checkpoints now store goal, constraints, next step, and dead ends, giving sessions that survive a compaction more context to resume from.

— Explains fields added but no exact surface name v5.13.64

04 **Date-scoped usage figures** IMPROVED 35

Usage figures are now dated by the day they occurred, so current-period metrics no longer include old session data.

— Clear behaviour change, no config surface given v5.13.64

WAS THIS USEFUL? 👍 👎

Charm Crush

Feedback ⓘ v0.100.0 NOTES ↗ CODE ↗ RANK

Crush is a terminal-based AI coding assistant that integrates LLMs, LSPs, and MCP servers into your development workflow.

> Crush v0.100.0 introduces fine-grained command permission controls, an unrestricted Sysadmin mode, in-terminal SSH/PGP authentication handling, and per-project configuration for routine commands.

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

01 Command permission management NEW

90

New `permissions safe`, `permissions block`, and `permissions unblock` commands let users mark commands as safe for yolo mode, block them entirely, or unblock previously restricted commands, giving fine-grained control over which commands the agent can run automatically.

Lock down infrastructure tools and whitelist safe build commands so yolo mode never touches your cluster.

```
$ permissions block kubectl terraform
permissions safe "go build" "cargo check"
permissions unblock curl
```

— Named subcommands with a runnable example v0.100.0

02 Sysadmin mode for unrestricted execution NEW

60

Sysadmin mode can be enabled via the command palette (`ctrl+p`) or the `shift+tab` shortcut to allow unrestricted command execution, including destructive operations.

— Exact key bindings given but no runnable command v0.100.0

THINNER COVERAGE BELOW

03 In-terminal SSH and PGP authentication IMPROVED

50

Crush now handles SSH password prompts and GPG/PGP pinentry directly inside the terminal, so private key passphrases and Yubikey interactions no longer require leaving the agent session.

— Describes behavior change but no concrete flag or command v0.100.0

04 **Per-project routine command configuration** NEW 20

Per-project configuration now lets projects define the commands Crush runs routinely.

— Bare description with no config key or mechanism given v0.100.0

└──▷ **ALSO FROM THESE RELEASES**

WAS THIS USEFUL? 👍 👎

Graphify Labs Graphify

Feedback ⓘ v0.9.103 NOTES ↗

RANK

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

> Graphify's release adds a `--depth` flag and persistent `query_budget` / `query_depth` settings for controlling query scope, alongside fixes to incremental updates, concurrent extraction, and the `export tree` command.

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

01 **Query depth and budget controls** NEW 90

The `query` command gains a `--depth` flag to cap how far a query traverses the graph, and `.graphifyrc` now supports persistent `query_budget` and `query_depth` keys so every agent invocation stays within token bounds without requiring per-run flags.

Narrow a query to a shallow subgraph to cut token spend when you only need the immediate neighbourhood of a concept.

```
$ graphify query "authentication flow" --depth 2
```

Set persistent query depth and budget limits for a project so every agent invocation stays within token bounds without requiring per-run flags.

```yaml
query_budget: 500
query_depth: 3
```

— Names exact flag and config keys with runnable examples v0.9.103

THINNER COVERAGE BELOW

02 **Capped directory output in export tree** IMPROVED 55

`export tree` output now caps the number of directory children shown while still reporting full totals.

— Names command and behaviour but no flag or limit value v0.9.103

03 **Thread-safe concurrent root extraction** IMPROVED 55

Concurrent library callers extracting different roots no longer clobber each other, because the ambient scan root is now held in a `ContextVar`.

– Explains mechanism but no user-facing action v0.9.103

04 **Pi project skill installation path** IMPROVED 55

The Pi project skill is now written to `.pi/skills`, and uninstall removes the global skill registration it wrote.

– Names exact path and uninstall behaviour v0.9.103

05 **AST cache schema bump to v39** BREAKING 55

The AST cache schema is bumped to 39; warm caches will re-run affected extractions on first use after upgrade.

– Clear migration note but no mitigation steps given v0.9.103

06 **Content-based change detection for incremental updates**

IMPROVED 20

Incremental updates are now made reliable via content-based change detection.

– Mentioned only in summary, no mechanism or scope given v0.9.103

└──▷ **BREAKING ON UPGRADE**

! The AST cache schema is bumped to 39; warm caches will re-run affected extractions on first use after upgrade.

WAS THIS USEFUL? 👍 👎

Tester Army e2e

Feedback ⓘ @e2e-dev/web@0.15.5 NOTES ↗

RANK

Next generation e2e testing framework for web and mobile apps.

> e2e's core 0.23.0 release adds a doctor command, an MCP run_tests tool, and an out-of-process engine protocol, while the web and mobile packages ship viewport resizing, binary response bodies, and tightened timeout, evaluate, and ARIA semantics alongside several breaking changes.

└──▷ **WHAT SHIPPED · 16 FEATURES** most completely described first ? what's the number?

01 e2e doctor command across core, web, and mobile NEW

93

`e2e doctor` checks that `e2e run` can start without actually running anything — covering the Node.js runtime, config, targets, engine checks, and resolved models — with `--json` printing a `doctor-1` report, `--check-model` calling each model once, and the command exiting 2 on failure. On mobile targets it additionally checks the agent-device, the platform toolchain, each configured device, an installed `bundleId`, and the `appPath` build, and on web targets it reports whether the browser a run will launch is already installed or will be downloaded on first run (both require e2e 0.22.1 or newer).

Pre-flight check that your config, engines, and model are all reachable before kicking off a CI run.

```
$ e2e doctor --check-model --json
```

Verify your mobile test environment is correctly configured before running a full suite — catches missing toolchains, misconfigured devices, and bad build paths early.

```
$ npx e2e doctor
```

— names flags, exit code, and per-target checks e2e@0.23.0 @e2e-dev/mobile@0.12.6 @e2e-dev/web@0.15.5

02 Viewport resizing for web targets and the built-in agent NEW

83

`session.resize` in the web engine exposes the agent's `resize_viewport` tool on web targets (requires e2e 0.22.1 or newer), and the built-in agent's new `resize_viewport` tool lets `agent.act` and `e2e explore` check layouts at

desktop, tablet, and phone widths; custom executors can call ctx.actions.resize({ width, height }) and the replay cache records and replays the resize.

— names API, tool, and replay-cache mechanism `e2e@0.23.0` `@e2e-dev/web@0.15.5`

---

03 **timeout: 0 semantics and clearer timeout error messages**

IMPROVED

83

Passing `timeout: 0` to locator actions, `waitFor`, matchers, `toHaveScreenshot`, `expect.poll`, `browser.goto`, `waitForResponse`, or `waitForDownload` removes the per-call deadline and waits until the test timeout instead; a negative, `NaN`, or infinite timeout value is now rejected as `INVALID_ARGUMENT` (and `expect.poll` with `timeout: 0` outside a test is `INVALID_CONFIG`). Timed-out fixture calls now name what they waited for in the error message, e.g. `operation timed out in browser.waitForResponse "foo.css"`.

Let a long-running navigation wait for the full test timeout instead of failing at a fixed per-call deadline.

typescript

```
await browser.goto('https://example.com/slow-page', { timeout: 0 });
```

— lists every affected API and the new error codes `@e2e-dev/web@0.15.5`

---

04 **Clearer test run reporting and diagnostics** IMPROVED

80

Failed-test blocks in the terminal and in `summary.md` now quote the first 3 app errors (console errors, uncaught exceptions, failed requests, 5xx responses) so an app error behind a locator failure shows without opening `trace.md`. The `list` reporter prints each test as it finishes when stdout isn't a terminal and ends every run with a `Failed` row per failed test (`file:line`, first error line, trace page). Filtered-locator failure messages now list how many candidates each filter step left, naming the step that narrowed too much or too little. The run summary's `Cache` line now counts steps whose recording was not saved (`step.cache.write` of `unconfirmed`, `no-change`, or `not-written`), e.g. `4 missed · 1 not saved`.

— names every reporter surface and example output `e2e@0.23.0`

---

05 **Binary response bodies in route.fulfill and route.fetch** NEW

78

`route.fulfill` now accepts a `Buffer` or `Uint8Array` body and sends the bytes unchanged, and `WebResponse` adds a body() method so route.fetch() and `waitForResponse` can pass binary response bodies through — useful for mocking image, PDF, or other binary API responses.

Intercept a request and reply with raw binary bytes rather than a JSON body — useful when mocking image, PDF, or other binary API responses.

```typescript
await route.fulfill({
  status: 200,
  headers: { 'content-type': 'image/png' },
  body: Buffer.from(pngBytes)
});
```

Inspect binary response bytes from a fetch intercepted by a route handler.

```typescript
const response = await route.fetch();
const bytes = await response.body();
await route.fulfill({ body: bytes });
```

— names the exact API and includes runnable examples `@e2e-dev/web@0.15.5`

06 **run_tests tool in e2e mcp** NEW 78

`e2e mcp` gains a `run_tests` tool so a coding agent can run a file, a single test (e.g. `tests/checkout.e2e.ts:12`), or the full suite and get back what `e2e run` printed, without needing a shell.

Let a coding agent run a single failing test and read the output without spawning a shell.

```
$ run_tests {files: ["tests/checkout.e2e.ts:12"]}
```

— names the MCP tool and shows a runnable call `e2e@0.23.0`

07 **Deeper, longer-running e2e explore sessions** IMPROVED 78

`e2e explore` now digs deeper once main flows are covered, returning with edge-case inputs, combined filters, reloads, and other accounts; `--max-steps` is raised to 100 and `--timeout` to 7200000 (2 hours).

Run a deep exploratory session for up to 2 hours to find edge-case UI failures across combined filters and multiple accounts.

```
$ e2e explore --max-steps 100 --timeout 7200000
```

— gives exact new flag defaults and a runnable command `e2e@0.23.0`

08 **app.baseUrl no longer has a trailing slash** BREAKING

75

`app.baseUrl` no longer ends with a trailing slash; a test that constructs URLs as `${app.baseUrl}api/users` or uses new URL('users', app.baseUrl) on a URL with a path must now add the slash itself.

— shows exact broken pattern and required fix `e2e@0.23.0` `@e2e-dev/web@0.15.5`

09 **browser.evaluate returns plain JSON only** BREAKING

73

`browser.evaluate` now returns plain JSON only, or `undefined` when the evaluated code returns nothing; returning a class instance such as DOMRect now fails with `INVALID_ARGUMENT` instead of arriving as a partial object, so callers must return rect.toJSON() or individual fields instead.

— names error code and the required migration `@e2e-dev/web@0.15.5`

10 **Startup and page-read performance improvements** IMPROVED

73

`e2e --version` and `--help` take about 25% less time, `e2e list` on a web project takes about 44% less time, and importing `@e2e-dev/web` takes about half as long (about 50%), achieved by deferring Playwright and other heavy dependencies until the command or session that needs them. Reading a 3000-node page now takes about 80ms instead of about 360ms by keeping elements in one in-page array per document instead of creating and releasing a handle for every node per observation.

— gives concrete before/after timings and mechanism `e2e@0.23.0` `@e2e-dev/web@0.15.5`

11 **Engine protocol for out-of-process engines** NEW

71

A new engine protocol — `connectEngine`, `serveEngine`, and transports in `e2e/engine` — lets an engine run in another process, on another machine, or in another language, communicating over LSP-style JSON-RPC.

— names the exported API surface `e2e@0.23.0`

12 **Hidden iframe elements now read as hidden** IMPROVED

65

Elements inside hidden `<iframe>` elements, including cross-origin frames, now read as hidden: isVisible() returns false, toBeHidden() passes, and `getByRole` with `{ visible: true }` skips them.

— names each API whose behavior changed `@e2e-dev/web@0.15.5`

13 **inputValue() and isChecked() now throw ACTION_FAILED** BREAKING

65

inputValue() on a node with no value (such as a heading or button) and isChecked() on a node with no checked state now throw `ACTION_FAILED` instead of returning `""` and `false`.

— names exact old vs new return behavior `e2e@0.23.0` `@e2e-dev/web@0.15.5`

---

14 **Mobile config validation and device-cloud session naming**

`IMPROVED`

65

mobile({ device }) now rejects a blank device name at config load instead of failing the run's boot, and device-cloud leases are now allocated and renewed under a named agent-device session, keeping BrowserStack, TestMu AI, and Limrun working with newer agent-device versions that refuse provider leases to unnamed clients.

— names config shape and three supported providers `@e2e-dev/mobile@0.12.6`

THINNER COVERAGE BELOW

15 **uncheck() on ARIA role=radio now clicks** `BREAKING`

50

uncheck() on an ARIA `role="radio"` now clicks the element, matching Playwright's behavior; a native radio input still refuses to uncheck.

— states before/after but no migration detail `e2e@0.23.0`

---

16 **SemanticNode.origin for tap positioning** `NEW`

40

Engines can now report `SemanticNode.origin` as the optional point that tap({ position }) measures from inside the node's border.

— single API name, no further mechanism given `@e2e-dev/web@0.15.5`

└──▷ **BREAKING ON UPGRADE**

! `app.baseUrl` no longer ends with a trailing slash; a test that constructs URLs as `${app.baseUrl}api/users` or uses new URL('users', app.baseUrl) on a URL with a path must now add the slash itself.

! inputValue() on a node with no value (such as a heading or button) and isChecked() on a node with no checked state now throw `ACTION_FAILED` instead of returning `""` and `false`.

! `browser.evaluate` now returns plain JSON only; a class instance such as `DOMRect` now fails with `INVALID_ARGUMENT` instead of arriving as a partial object — callers must return rect.toJSON() or individual fields instead.

! A negative, `NaN`, or infinite `timeout` value passed to any locator action, `waitFor`, matcher, or `browser.*` call is now `INVALID_ARGUMENT`; `expect.poll` with `timeout: 0` outside a test is `INVALID_CONFIG`.

! inputValue() on a node with no value (such as a heading or button) now throws `ACTION_FAILED` instead of returning `""`.

- ! isChecked() on a node with no checked state now throws `ACTION_FAILED` instead of returning `false`.
- ! uncheck() on an ARIA `role="radio"` now clicks the element (as Playwright does); a native radio still refuses.
- ! `app.baseUrl` no longer ends in a trailing slash — a test that depended on it, such as `${app.baseUrl}api/users` or new URL('users', app.baseUrl) on a URL with a path, must add the slash itself.

WAS THIS USEFUL? 👍 👎

flow-next

Feedback ⓘ flow-next-v9.6.0 NOTES ↗ RANK

flow-next is a workflow plugin that takes a bug, idea or ticket to a verified pull request with specs, cross-model review, live QA, and receipts

> flow-next's headline addition this window is full GitLab merge request support, letting `flow --auto --until=merge` run the entire open-review-merge loop end to end, alongside smaller additions for spec-tracker sync, draft MR control, and automation introspection.

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

01 GitLab merge request support end-to-end

BREAKING 91

flow-next can now run the full open-review-merge loop on GitLab via `flow --auto --until=merge`, which lands on a GitLab origin instead of stopping after `make-pr`. The `land` command squash-merges only when the pipeline for the current head has passed and GitLab's merge check allows it, with the head pinned so a late push fails the merge. Authenticate by logging in with `glab` using the `api` scope — on a GitLab origin, `make-pr`, `resolve-pr`, and `land` now act through `glab` instead of stopping, so this `glab auth login` is required before upgrading.

Run unattended from spec to merged GitLab MR — no stops unless CI fails or a conflict blocks land.

```
$ flow --auto --until=merge
```

Run the entire GitLab merge request lifecycle — open MR, resolve discussions, and merge — without manual steps.

```
$ flow --auto --until=merge
```

— Names exact command, conditions, and required pre-upgrade auth step. flow-next-v9.6.0

02 `tracker.specWins` syncs tracker issue to spec

NEW 80

Enable `tracker.specWins: true` in `.flow/config.yaml` to make the tracker issue body follow the spec — written at first claim and again at make-pr — so the spec is always the record and no edit made directly on the tracker survives make-pr or first claim.

Keep the tracker issue body in sync with the spec so no edit made directly on the tracker survives make-pr or first claim.

```yaml
tracker:
  specWins: true
```

— Exact config key, file path, and behaviour given. flow-next-v9.6.0

03 **Draft merge request flags for GitLab `make-pr`** NEW

75

Use `--draft`, `--ready`, and `--update` flags with `make-pr` to control draft merge request state on GitLab, mirroring the existing GitHub behaviour.

— Named flags with clear GitHub-parity behaviour, no example run. flow-next-v9.6.0

THINNER COVERAGE BELOW

04 **`flowctl pilot snapshot` reports GitLab MR state** IMPROVED

55

`flowctl pilot snapshot` now reports the origin's `forge` and an open merge request's readiness, so automation can inspect GitLab MR state.

— Names the command and fields but no concrete output shown. flow-next-v9.6.0

05 **Per-child review ceremony control in goal spec** NEW

40

flow-next lets you control how much review ceremony each child task gets — from straight commits to a full pull request and review per child — by specifying it in the goal spec.

— Describes capability in prose without naming the spec field. product docs

└──▷ **BREAKING ON UPGRADE**

! On a GitLab origin, `make-pr`, `resolve-pr`, and `land` now act through `glab` instead of stopping, so a `glab auth login` with the `api` scope is required before upgrade.

WAS THIS USEFUL? 👍 👎

Amazon Kiro Crew

Feedback ⓘ v0.12.0 NOTES ↗ RANK

Kiro Crew is an open-source workspace for coordinating AI coding agents.

> Kiro Crew's v0.12.0 release opened Bots to all users, added live browser takeover and new Strands and GitHub Copilot CLI harnesses, and let agents connect directly to Google Workspace and Amazon Bedrock Knowledge Bases, alongside new Slack code channels, a batch-write safety guard, expanded cron scheduling, and several CLI additions.

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

01 Cron job scheduling options NEW

90

`kirocrew cron add --alert-after N` rings the bell only for failures after N runs and shows Partial when the security gate partly refuses a run; `kirocrew cron add --at` accepts natural-language one-shot times such as in 2 days or in 1 week ; and `kirocrew cron add --project-path <dir>` runs the job's agent session in a specific directory so the repo's steering files load.

Schedule a one-shot agent job to run in two days, scoped to a specific repo directory so its steering files are picked up.

```
$ kirocrew cron add --at 'in 2 days' --project-path
/home/user/myrepo --alert-after 1 'Run nightly audit'
```

— Three named flags with exact behaviour and a runnable example v0.12.0

02 Slack code channels for multi-step work NEW

83

Enable Slack code channels with `slack.code_channels` (off by default) so a thread mention describing multi-step work opens a code channel backed by a dashboard session, which can be stopped via its button or `!stop` ; requires your workspace's Slack code channels feature, an updated app manifest, and a reinstall.

— Config key, trigger behaviour, and stop command all named v0.12.0

03 Persistent MCP sign-ins across sessions NEW

81

Remote MCP sign-ins such as Atlassian, Notion, and Figma are now kept alive across sessions by default on the KAS backend; disable this persistence with `agent.kas_persist_mcp_sign_in` .

Silence persistent MCP re-authentication prompts for Atlassian, Notion, and Figma by keeping sign-ins alive across sessions.

```yaml
agent.kas_persist_mcp_sign_in: true
```

— Exact config key and default behaviour with runnable example v0.12.0

04 **Bedrock Knowledge Base registration and live query** NEW

80

Register an existing Amazon Bedrock Knowledge Base under `Capabilities > Knowledge > Sources > 'Add Knowledge Base'` to query it live and merge results into knowledge search; this requires the optional `bedrock` extra, and each search is billed by AWS.

— Exact UI path and extra named, but no further mechanism given v0.12.0

05 **Update installed app from source directory** IMPROVED

80

`kirocrew app update <name> --source <dir>` updates an installed app from another directory and remembers that directory as the app's source for future updates.

Update a deployed app from a local source directory, re-registering its agents and skills in place.

```
$ kirocrew app update my-security-app --source ./apps/my-security-app
```

— Exact command and flag with runnable example v0.12.0

06 **Batch write guard for parallel tool calls** NEW

78

Setting `hooks.batch_write_guard` (off by default) makes Kiro Crew refuse writes sent in the same parallel batch as a read, preventing speculative writes based on guessed arguments.

Prevent an agent from making speculative writes based on guessed arguments when it has also issued a read in the same parallel batch.

```yaml
hooks.batch_write_guard: true
```

— Named config key with clear mechanism and example v0.12.0

07 **Open sessions in desktop app via deep link** NEW

63

Open a session in the desktop app via `kirocrew://chat?sid=<key>` or with `kirocrew open`.

— Named deep link and command but no further mechanism v0.12.0

08 **Google Workspace connections for agents** NEW 60

Agents can connect to Gmail, Google Drive, Calendar, Sheets, Docs, and Slides via the Connections page using your own Google OAuth app.

— Named apps and UI location, but setup mechanism not detailed v0.12.0

THINNER COVERAGE BELOW

09 **Environment version check command** NEW 52

`kirocrew doctor` checks installed and needed engine versions.

— Single named command with minimal elaboration v0.12.0

10 **Strands and GitHub Copilot CLI harnesses** NEW 25

This release added Strands and GitHub Copilot CLI harnesses for running agents, though no further detail on configuration or behaviour was given.

— Named only in summary, no mechanism or usage detail v0.12.0

11 **Live browser takeover** NEW 16

Kiro Crew v0.12.0 added live browser takeover; no further detail on how it works was provided.

— Named only in release summary, no mechanism described v0.12.0

12 **Bots available to all users** IMPROVED 13

Bots are now available to all users of Kiro Crew, per the release summary; no additional detail was given.

— Only named in summary, no scope or mechanism given v0.12.0

WAS THIS USEFUL? 👍 👎

pstack-claude

Feedback ⓘ v0.9.130 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 adds native plugin installs for four new AI coding platforms—Kimi Code CLI, Grok Build, OMP, and OpenCode 2—while improving deslop's branch-diffing and check-reverting logic and extending Codex rollout history support; the new OpenCode 2 plugin replaces the old symlink-based install and drops OpenCode 1 support.

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

01 Native plugins for Kimi Code CLI, Grok Build, OMP, OpenCode 2

NEW 95

pstack can now be installed as a native plugin on four more runtimes. Kimi Code CLI: `/plugins install https://github.com/michael-denyer/pstack-claude` then `/reload`, skills invoked as `/skill:pstack:poteto-mode`; `kimi.plugin.json` wires the shared skills tree and routing instruction, and `kimi-tools.md` maps Claude tools to Kimi's `Agent`, `coder`, `explore`, `plan`, and `CronCreate`. Grok Build: install from a local checkout with `grok plugin install "$PWD/plugins/pstack" --trust` and `grok plugin enable pstack`; `.grok-plugin/plugin.json` lists the shared skills plus a Grok-only `pstack-routing` skill, and `grok.md` maps Claude tools to Grok's including `spawn_subagent`. OMP: `omp plugin marketplace add michael-denyer/pstack-claude` then `omp plugin install --scope user pstack@pstack-claude`; `plugins/pstack/package.json` and `omp/index.mjs` register skills as `pstack:<name>`, and setting `session hook: off` in `~/.omp/pstack-models.md` disables routing for every OMP profile. OpenCode 2: add `plugins/pstack/opencode` to the `plugins` array in `~/.config/opencode/opencode.jsonc`, or run `opencode plugin add 'github:michael-denyer/pstack-claude#main'`; `opencode/index.js` registers each skill as `pstack-<name>` and injects routing at every model step including after compaction, and `session hook: off` in `${XDG_CONFIG_HOME:-~/.config}/opencode/pstack-models.md` turns it off.

— Exact install commands and config files for all four platforms given v0.9.130

02 Codex rollout history mapping for recall and show-me-your-work

IMPROVED 80

`codex-tools.md` now maps Codex rollouts under `${CODEX_HOME:-~/.codex}/sessions/` and `archived_sessions/`, including zstd-compressed `.jsonl.zst` files read with `zstd -dc`, filtered to rollouts whose `cwd` matches the

workspace or a worktree; `recall` and `show-me-your-work` now read Codex history using this mapping.

— Exact paths, compression format, and consuming commands all named v0.9.130

---

03 **OpenCode 2 plugin replaces legacy symlink install** BREAKING

65

The OpenCode 2 plugin replaces the previous `~/.agents/skills` install method; existing symlinks pointing at a pstack checkout in `~/.agents/skills` must be removed before enabling the new plugin, and OpenCode 1 is no longer supported.

— Names exact migration step but omits further rationale v0.9.130

---

04 **deslop diffs against PR base branch and reverts failing edits**

IMPROVED

65

`deslop` now diffs against the PR's base branch (or the branch it was cut from, or the default branch) rather than always against `main`, and reruns checks after each edit to revert any change whose check fails.

— Describes new branch logic and auto-revert but no command example v0.9.130

└──▷ **BREAKING ON UPGRADE**

! The OpenCode 2 plugin replaces the previous `~/.agents/skills` install method; symlinks pointing at your pstack checkout in `~/.agents/skills` must be removed before enabling the new plugin, and OpenCode 1 is no longer supported.

WAS THIS USEFUL?  

The Open Engine Zeroshot

Feedback ⓘ v11.1.0 NOTES ↗

RANK

Zeroshot orchestrates executor-verifier agent loops that implement and validate software changes while logging workflow state to SQLite.

> Zeroshot's v11.1.0 release adds a local `script` runtime binding for graph steps, a built-in `code-review` template, commit identity control for hosted runs, and opt-outs for automatic skill installation and workspace checkpoints.

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

01 Opt-out for workspace checkpoints in runtime plans NEW

86

Setting `workspaceCheckpoints: false` in a runtime plan disables workspace checkpoints, skipping Restic setup, workspace snapshots and snapshot cleanup while retaining durable run history, outputs, logs and usage.

Speed up local runs by disabling workspace snapshots in a runtime plan while keeping durable history and logs.

json

```
{
  "workspaceCheckpoints": false
}
```

— Names config key and exact before/after behaviour of what's skipped vs retained v11.1.0

02 Custom commit identity for hosted runs NEW

86

Hosted runs can set the author and committer of delivery commits via `TargetRunRequest.commitIdentity` (`{name, email}`) on submission or `RunResumeParams.commitIdentity` for a resumed successor; invalid or absent values fall back to `Zeroshot <delivery@zeroshot.invalid>`, and values longer than 256 bytes or containing control characters, `<`, or `>` are rejected with `400 request.invalid`.

— Exact field names, fallback identity, and error response code given v11.1.0

03 Built-in code-review template NEW

79

A `code-review` built-in template inspects code and existing tests without installing dependencies or starting applications, services or browsers; it accepts `--task-from <PR URL>` or `{url, context, instructions?}` input and delivers one comment.

— Names exact flag and input schema plus behavioural scope v11.1.0

04 **Local script runtime binding for graph steps** NEW 70

A local `script` runtime binding can be added to typed graph steps, supplying an executable and literal arguments that exchange bounded JSON over stdin and stdout, validate results against the step's output schema, and record execution timings with zero native model tokens.

— Mechanism well explained but no concrete config example or command shown v11.1.0

05 **Opt-out for automatic skill installation** NEW 68

Set `ZEROSHOT_SKIP_SKILLS=1` to skip automatic skill installation during `npm install` and self-update.

Opt out of automatic skill installation in a CI environment where you manage skills yourself.

```
$ ZEROSHOT_SKIP_SKILLS=1 npm install -g @the-open-engine-company/zeroshot
```

— Exact env var and command given, but skill install mechanism itself unexplained v11.1.0

THINNER COVERAGE BELOW

06 **Minimal environment snapshot for embedded hosts** NEW 40

An embedded host can save an environment snapshot acknowledged with only its identity, name, definition, and revision, keeping host-owned settings out of Zeroshot state.

— Describes scope but no API name, field list, or config surface given v11.1.0

WAS THIS USEFUL? 👍 👎

Alibaba OpenCodeReview

Feedback ⓘ v1.13.6 NOTES ↗

Secure, fast, efficient, battle-tested at Alibaba's scale. Hybrid architecture code review tool: deterministic pipelines + LLM Agent, precise line-level comments, built-in multi-language ruleset (NPE, thread-safety, XSS, SQL injection), OpenAI & Anthropic compatible.

└──▷ GET THIS VERSION ▾

Homebrew

```
$ brew install open-code-review
# already installed? upgrade to this version:
$ brew upgrade open-code-review
```

npm

```
$ npm install -g @alibaba-group/open-code-review@1.13.6
```

git

```
$ git clone --branch v1.13.6 https://github.com/alibaba/open-code-review.git
# already have the repo? check out this version:
$ git checkout v1.13.6
```

Homebrew lists this version · signed build provenance · checked 6 Nov Wrong command?

NEW SUPPORTED PLATFORM Starlark file review support

NEW DETECTION C# review rules

WAS THIS USEFUL? 👍 👎

Sourcegraph Amp

Feedback ⓘ GitLab Integration (2026-11-05)

BLOG ↗

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

Amp adds a GitLab integration with merge request automation, CI subscription, per-issue orb spawners, and service account support for workspace admins.

NEW INTEGRATION Connect GitLab accounts in Settings → Integrations & MCP to start projects from GitLab repositories on gitlab.com or self-managed instances.

Manage event subscriptions, pause or delete spawners, edit a spawner, and view orbs each spawner started under `Settings → Integrations & MCP → GitLab` in the Event Subscriptions section.

Orbs now clone, push, and run with `glab` signed in so the agent can open merge requests, comment on them, and read issues and pipelines.

Subscribe Amp to comments and CI for a merge request it opens so that reviewer requests or failed pipelines land in the originating thread with full context.

Automatically start a new orb for every new merge request or issue in a project using an orb spawner — each orb reads the item and has a summary ready when you open the thread.

+1 more

WAS THIS USEFUL? 👍 👎

Amazon Kiro

Feedback ⓘ

Models: Generally Available: Claude Haiku 5.5, Kimi K3 (2026-11-04) BLOG ↗

RANK

Kiro is an agentic development environment that uses spec-driven workflows to plan, build, and maintain software.

> Kiro promoted Claude Haiku 5.5 and Kimi K3 to general availability across its IDE, CLI, Crew, and Web surfaces, with an accompanying change to indemnity coverage for Claude Haiku 5.5.

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

01 Claude Haiku 5.5 and Kimi K3 reach general availability NEW

75

Claude Haiku 5.5 and Kimi K3 are now generally available across the Kiro IDE, CLI, Crew, and Web, with their Experimental labels removed. Reload the model list by restarting the IDE or CLI, running `kirocrew restart` for Crew, or refreshing Web in the browser. Claude Haiku 5.5 now qualifies for Kiro's standard indemnity coverage (the Kiro Excluded Models exception no longer applies to it), while Kimi K3 remains a Kiro Excluded Model.

Reload the available model list in Kiro Crew after the GA promotion so agents can select Claude Haiku 5.5 or Kimi K3.

```
$ kirocrew restart
```

— Names exact products, labels, command, and indemnity status change.

Models: Generally Available: Claude Hai…

WAS THIS USEFUL? 👍 👎

◆ Model Training & Fine-Tuning

fizgig

Feedback ⓘ v9.0.6 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's biggest window update is batch training support — up to 8 pictures per step on most models with automatic VRAM/speed estimation — alongside a new Identity mode for LoRA extraction and training, an extension of per-image loss watching to batched runs, and a breaking rename of Klein 9B's VAE file.

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

01 Batch size training across fine-tunes and most models NEW

90

Full fine-tunes now support Batch size, training several pictures per step; Fizgig sizes windows to accommodate the extra pictures, the fine-tune card's plan line shows the result for the current card, and Move to… selects the largest batch that keeps the window plan intact. This was extended to train up to 8 pictures per step by setting Batch size (under Other Options on the Training tab) to values 2–8 on all models except ACE-Step, with the Training tab showing estimated VRAM use and likely speed at the chosen batch versus batch 1. A new Train several pictures a step there option in Move to… moves a run to a larger remote pod and picks the largest batch the pod's card can fit at the run's precision without extra block swaps.

— Names exact UI location, flag values 2–8, and Move to… behaviour. v9.0.6 v9.0.5

02 Identity mode for LoRA extraction and MOSS-Image training

NEW

83

Identity mode in Extract (Fine-tune to LoRA or From a LoRA) gives MOSS-Image blocks 24–35 and Qwen Image 2.1 blocks 10–14 four times the chosen rank while all other blocks keep the base rank, preserving character likeness with a smaller file than raising rank globally; extracted LoRAs are saved with `_id` in the filename. MOSS-Image also gained Fast Identity Mode on the Training tab, plus Studio's identity and look colours with Identity only, Look only and Identity ×0.5 presets.

— Names exact blocks, rank multiplier and filename convention. v9.0.6

03 Klein 9B VAE filename change BREAKING

83

Klein 9B's VAE is now saved as `flux2-vae.safetensors` instead of `ae.safetensors`; on upgrade Fizgig moves existing VAE rows automatically, but rows

pointing at the wrong file are cleared and must be re-downloaded via Download models for me or Start.

— Exact filenames and clear migration/recovery steps given. v9.0.5

---

04 **Per-image loss watch works with batched training** IMPROVED

65

The per-image loss watch — Detect problem images, Per-image adaptive LR, look-outlier warm-up and Auto-recaption — now works at any Batch size above 1, weighing, excluding and recording each picture in a batch independently.

— Names the four features but no UI path given. v9.0.6

└──▷ **BREAKING ON UPGRADE**

! Klein 9B's VAE is now saved as `flux2-vae.safetensors` instead of `ae.safetensors`; on upgrade Fizgig moves existing VAE rows automatically, but rows pointing at the wrong file are cleared and must be re-downloaded via Download models for me or Start.

WAS THIS USEFUL? 👍 👎

Thinking Machines Lab Tinker

Feedback ⓘ SDK 0.38.0

NOTES ↗ RANK

Tinker provides a managed API for training and sampling language models while users control their training algorithms and data.

> Tinker SDK 0.38.0 unifies all client methods around a new APIFuture return type with sync and async bridges, while deprecating the older async method and Future-based patterns it replaces.

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

01 APIFuture unified return type across client methods NEW

95

All client methods now return an `APIFuture`: call .result() in sync code or `await` it directly in async code. New .future() and .coroutine() helpers convert any `APIFuture` into a standard `concurrent.futures.Future` (for use with `concurrent.futures.wait`, `as_completed`, or `add_done_callback`) or into a coroutine (for use with `asyncio.create_task`). sample() and compute_logprobs() now return an `APIFuture` instead of a `concurrent.futures.Future`, as do the client-creation methods `create_lora_training_client`, `create_training_client_from_state`, `create_training_client_from_state_with_optimizer`, `create_sampling_client`, and `save_weights_and_get_sampling_client`, and all `RestClient` and `AdminClient` methods.

Fan out multiple sample calls concurrently and collect results as they complete, using the new .future() bridge.

python

```python
from concurrent.futures import as_completed

futures = [client.sample(prompt=p).future() for p in prompts]
for f in as_completed(futures):
    print(f.result())
```

Schedule a sample call as a background asyncio task using the new .coroutine() helper.

```python
import asyncio

task =
asyncio.create_task(client.sample(prompt='Hello').coroutine())
result = await task
```

— Named methods, helpers, and runnable code examples given. SDK 0.38.0

02 **Deprecation of legacy async method and Future patterns**

85

BREAKING

The `_async` methods on `ServiceClient`, `TrainingClient`, and `SamplingClient` are deprecated and now emit a `FutureWarning`, to be removed in a subsequent release. Calling `concurrent.futures.Future` methods such as .done() or .add_done_callback() directly on objects returned by sample() or compute_logprobs() now emits a `FutureWarning` — callers must call .future() first. Using the object returned by client-creation methods (`create_lora_training_client`, `create_training_client_from_state`, `create_training_client_from_state_with_optimizer`, `create_sampling_client`, `save_weights_and_get_sampling_client`) directly as the client is also deprecated.

— Clear migration note but no single runnable fix shown. SDK 0.38.0

THINNER COVERAGE BELOW

03 **Faster get_info() and get_base_model() calls** IMPROVED

55

TrainingClient.get_info() and SamplingClient.get_base_model() now return immediately instead of making a network call.

— States the change but no mechanism detail. SDK 0.38.0

04 **get_server_capabilities() moved to RestClient** DEPRECATED

50

get_server_capabilities() moves to `RestClient`; ServiceClient.get_server_capabilities() is deprecated in its favor.

— Names old and new locations but no further detail. SDK 0.38.0

└──▷ BREAKING ON UPGRADE

! The `_async` methods on `ServiceClient`, `TrainingClient`, and `SamplingClient` are deprecated and emit a `FutureWarning`; they will be removed in a subsequent release.

! Calling `concurrent.futures.Future` methods such as .done() or .add_done_callback() directly on the object returned by sample() or compute_logprobs() now emits a `FutureWarning` ; callers must switch to .future() before calling those methods.

! Using the object returned by client-creation methods (`create_lora_training_client` , `create_training_client_from_state` , `create_training_client_from_state_with_optimizer` , `create_sampling_client` , `save_weights_and_get_sampling_client`) directly as the client is deprecated.

! ServiceClient.get_server_capabilities() is deprecated in favor of RestClient.get_server_capabilities().

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

Brig Feedback ⓘ v0.5.2 NOTES ↗ RANK

Brig runs coding agents in a microVM sandbox, made up of the brig CLI and the brigd session daemon.

> Brig adds GCP route support, letting sandboxed agents authenticate to Google Cloud services using the host's gcloud credentials.

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

01 **GCP route support via host gcloud credentials** NEW

67

Brig adds a `gcp` policy preset listing GCP services, so a named egress policy can be applied to a GCP-bound sandbox without writing service URLs by hand. The broker routes GCP traffic to each service's host and injects a minted Google token on the route, authenticating sandboxed agents as the host's gcloud identity, and Brig mints and caches these Google tokens per sandbox and route so each sandbox gets its own credential without re-authenticating.

— Names the `gcp` preset and token mechanism but no exact command shown v0.5.2

WAS THIS USEFUL? 👍 👎

CopilotKit OpenBot

Feedback ⓘ v0.3.1 NOTES ↗ RANK

OpenBot runs self-hosted AI coworkers that each get their own browser, files, and gated tools, with every action decided and recorded.

> OpenBot's two releases this window add a live terminal and persistent shell to the Computer panel, a Recorder and Learning space that replaces browser-workflow teaching, and shared per-person computers, alongside a wide set of breaking changes to model defaults, broker provider selection, and approval requirements for sensitive Bot actions.

└──▷ **WHAT SHIPPED · 20 FEATURES** most completely described first ? what's the number?

01 **Computer resource management config flags** BREAKING

93

Hosted computer browsers now default to headed Chromium on a virtual display instead of headless, using more memory; set `COMPUTER_BROWSER_MODE=headless` (Helm: `computers.browserMode: headless`) to restore headless mode. Browsers now close automatically once host memory crosses 85%, configurable with `COMPUTER_MEMORY_HIGH_PERCENT=0` to disable (Helm: `computers.memoryHighPercent`). Idle tabs close after ten minutes by default, configurable with `COMPUTER_TAB_IDLE_MS=0` to disable (Helm: `computers.tabIdleMs`). Automatic memory source defaults for connected apps can be turned off with `OPENBOT_MEMORY_SOURCE_DEFAULTS=false` (Helm: `config.memory.sourceDefaults: false`).

Run a hosted computer's browser in headless mode to conserve memory when a full display is not needed.

```env
COMPUTER_BROWSER_MODE=headless
```

Prevent OpenBot from auto-assigning a Learning container and opt out of automatic memory source defaults in a Helm deployment.

```yaml
config:
  learning:
    autoAssign: false
  memory:
    sourceDefaults: false
```

— Every flag, Helm key, default and limit named verbatim. v0.3.0

02 **Live terminal tab and persistent shell per Bot** NEW

88

The Computer panel gains a Terminal tab alongside the Screen, showing a Bot's shell live: commands appear as typed, output streams in colour, and Take control hands you the keyboard while the Bot waits. Commands also render in the conversation and in Activity as terminal blocks with the prompt, command, last lines of output with Show all N lines, and an end state (Exit 0 , Exit 1 , Running , Blocked , Declined , or Done by you); clicking one opens the Computer panel on that command's terminal. Bot commands now run in one persistent interactive bash session per Bot per person instead of a fresh `bash -c` each time, so `cd` , exported variables, and background processes (`&`) carry over between commands, and `COMPUTER_SHELL_ENV` lets you

pass custom `TERM`, `PAGER`, `GIT_PAGER`, and `GIT_TERMINAL_PROMPT` values into that persistent shell.

Pass a custom pager and suppress Git prompts in every Bot shell without modifying the container image.

```env
COMPUTER_SHELL_ENV=PAGER=less GIT_PAGER=less
GIT_TERMINAL_PROMPT=0
```

— Full mechanism and config key given, but no UI screenshot. `v0.3.1`

---

03 **Shared per-person computers with isolation control** BREAKING

87

Bots on shared computers now share one computer per person instead of running per-Bot containers; the supervisor stops the old per-Bot containers on start and leaves their volumes for manual removal. Isolation is controlled by the new `agent_profiles.isolated` column (migration `0069_isolated_bots`), and under `computers.mode: shared` each person's files per Bot now live at `computers/<person id>~<Bot id>` on the workspace volume — files previously at the workspace root must be moved there to stay visible. Helm's `computers.browserIdleMs` and `computers.maxBrowsers` now also take effect on `sandbox` computers.

— Names the migration, column, and the exact new file path. `v0.3.0`

---

04 **Default model selection and BOT_MODEL behavior** BREAKING

83

When no model is configured, built-in Bots now answer on `BOT_MODEL` if the provider lists it, otherwise on the middle model of the provider's current line — as of 2026-11-05 that is `gpt-6.1-terra`, `claude-sonnet-5-5`, `gemini-3.8-flash`, or `grok-4.7`. `BOT_MODEL` can still be set to override the Anthropic default (now `claude-sonnet-5-5`) or the OpenAI router model, and an existing value is left unchanged; but nothing writes `BOT_MODEL` on an installation's behalf any more — `docker-compose.yml` now leaves it empty when unset (previously `gpt-5.5`) and OpenBot Desktop no longer writes `BOT_MODEL=claude-sonnet-5-5` for an Anthropic key, removing any it previously wrote.

— Exact model names and env var given, but spans two releases. `v0.3.1` `v0.3.0`

---

05 **Recorder and Learning space replace browser workflow teaching**

NEW

83

OpenBot adds a Recorder and a per-person Learning space, backed by six new tables (`recorder_connect_codes`, `recorder_connections`, `recorder_trajectory_bindings`, `learning_person_spaces`, `learning_skill_sharing`, `learning_open_skills`), activated when someone uses Teach a Bot, connects a browser, or visits Settings → Learning. The Teach a browser workflow feature and its routes under `/api/demonstrations` are removed — along with the instant Skill draft, Save and use with this Bot, and Run on a schedule — with recording now done in Teach a Bot or in a Bot's browser; prior recordings remain in the database but are unread.

Prevent OpenBot from auto-assigning a Learning container and opt out of automatic memory source defaults in a Helm deployment.

```yaml
config:
  learning:
    autoAssign: false
  memory:
    sourceDefaults: false
```

— Names every table and the replaced routes; no UI walkthrough. v0.3.0

06 **Configurable model capability list source** NEW 81

`OPENBOT_MODEL_CAPABILITIES_URL` can point at a mirror of models.dev, or be set to `off` to skip external model-list fetching entirely; the server now reads the provider's model list at startup and every six hours.

Mirror the models.dev capability list internally so the server never fetches it from the public internet.

```text
OPENBOT_MODEL_CAPABILITIES_URL=https://mirrors.internal/models-dev/api.json
```

Disable the external model-list fetch entirely and lock the server to whatever model is configured.

```env
OPENBOT_MODEL_CAPABILITIES_URL=off
```

— Named env var with exact values and refresh interval. v0.3.1

07 **Hostname allowlisting and header forwarding changes** BREAKING

80

Single-user OpenBot now refuses requests addressed to unrecognised hostnames and blocks other machines by default; set `OPENBOT_PUBLIC_URL`, `OPENBOT_APP_URL`, or `TRUSTED_ORIGINS` to allow them, and a value with no scheme (e.g. `example.com:3010`) now stops the server from starting instead of only warning. Bot endpoints also no longer forward Authorization or `x-*` headers from the person's request; remote Bots that read these must be given them as the Bot's own configured headers.

— Names every env var and the exact header behaviour change. v0.3.0

---

08 **Admin-controlled model allowlist in Providers** BREAKING

78

Admin → Providers now lets team admins control which models people may use via an explicit allowlist under Allowed models, replacing the Let people choose a model switch. Setting `OPENBOT_SINGLE_USER=true` offers every model the provider lists with no allowlist, and in that mode Admin → Providers no longer shows an allowlist or switch; with Only allowed models on, the default model must be explicitly allowed or unpicked runs are refused, and a pick that is no longer allowed or listed silently falls back to the default.

— Names the UI path, env var, and fallback behaviour. v0.3.1

---

09 **Gemini request path pin for Python Bot harnesses** BREAKING

78

The Python LangGraph and Pydantic AI Bots now pin `google-genai` at `>=2.34,<3`; images built from source after `google-genai` 3.0.0 (2026-11-06) had installed 3.x and sent Gemini requests to `/v1` instead of `/v1beta` — rebuilding from this release restores `/v1beta`.

— Exact version pin and endpoint paths named; fix is a rebuild. v0.3.1

---

10 **Explicit broker provider selection and single-user isolation**

78

BREAKING

With both `COMPOSIO_API_KEY` and `ARCADE_API_KEY` set, the server now refuses to start unless `OPENBOT_BROKER_PROVIDER=composio` (or the Arcade key is unset) declares the choice explicitly. Single-user deployments on Composio now get their own user id at the broker on first start (migration `0074_single_user_identity`), isolating broker connections from shared keys; existing connections recorded under `dev-local-user` are removed from OpenBot and each app must be reconnected once after upgrading.

Declare Composio as the broker provider and name a custom router model when both broker keys are present.

```env
OPENBOT_BROKER_PROVIDER=composio
ROUTER_MODEL=claude-haiku-5-5
```

— Names the env var, migration id, and the reconnect requirement. v0.3.0

11 **OpenAI Responses vs Chat Completions API selection** BREAKING

75

Behind a custom `OPENAI_BASE_URL` that is not OpenAI or Azure, built-in Bots now use Chat Completions instead of the Responses API by default; set `COPILOTKIT_OPENAI_API=responses` to force the Responses API for a gateway that only serves that API.

— Names the exact env var and the default behaviour change. v0.3.0

12 **Custom router model for untagged messages** NEW

70

Name a custom router model for untagged home-screen messages with `ROUTER_MODEL`; without it, routing uses `claude-haiku-5-5` on Anthropic or `gpt-6-luna` on OpenAI.

Declare Composio as the broker provider and name a custom router model when both broker keys are present.

```env
OPENBOT_BROKER_PROVIDER=composio
ROUTER_MODEL=claude-haiku-5-5
```

— Names the env var and both default models. v0.3.0

13 **Approval required for sensitive and refund actions** BREAKING

67

Bot actions that delete, send, share, post, or schedule sends now require approval every time — any previously saved Always allow for these actions, including grants made on private-share cards, no longer applies. Built-in safety rules now hand refunds to the person instead of letting the Bot issue them automatically; Bots that previously issued refunds (e.g. via `SHOPIFY_CREATE_REFUND`) now return each one to the user for approval.

— Names the affected action types and one tool call example. v0.3.0

14 **Telemetry opt-out flags for Desktop** NEW

60

Launch OpenBot Desktop with `DO_NOT_TRACK=1` or `COPILOTKIT_TELEMETRY_DISABLED=1` to opt a deployment out of product usage telemetry.

— Exact env vars given, but no detail on what telemetry covers. v0.3.0

THINNER COVERAGE BELOW

15 **Enterprise admin settings redistributed across pages** BREAKING

57

Admin → Enterprise controls is removed; its settings are now distributed across Permissions, Providers, Identity providers, Computers, Plugins, Automatic learning, and Audit pages, and `/admin/enterprise` redirects to the Admin overview.

— Names the destination pages but not what moved where. v0.3.0

16 **Forced proactive check-in defaults** NEW

55

Setting `OPENBOT_PROACTIVE_DEFAULTS=true` forces proactive check-in defaults on; check-ins are otherwise not switched on automatically.

— Names the flag but not what a check-in does. v0.3.0

17 **Desktop install auto-update on Start with rollback** IMPROVED

50

OpenBot Desktop now moves an installation to the app's current release on Start through the same path as Runtime updates, performing a database backup first and automatically rolling back if the move fails after stopping.

— Describes mechanism but no manual trigger or flag. v0.3.1

18 **Routines made permanent and non-removable** BREAKING

50

Routines is now always on in every edition: the Routines app and its tools are added to every Bot automatically and cannot be removed (attempts return `403`), and an MCP allowlist cannot exclude it.

— Names the HTTP status but not what Routines does. v0.3.0

19 **Composio key auto-provisioning for Desktop** IMPROVED

45

OpenBot Desktop installs signed in to CopilotKit now automatically receive a Composio key in the app's protected store (never written to `.env`) at their next Start, restoring app availability that had been lost in 0.3.0.

— Explains the fix and storage location but nothing more to act on. v0.3.1

20 **Hover-accurate pointer input on live screen** IMPROVED

35

The live screen now sends hover pointer positions so hovering registers as a hover rather than a drag, and input prioritisation ensures clicks and keys always land in order and are never merged or dropped.

— Describes the fix but gives no config or test steps. v0.3.1

└──▷ BREAKING ON UPGRADE

! Bot commands now run in a single persistent bash session per person per Bot rather than a fresh `bash -c` each time — scripts that relied on a clean environment for every command will see state from prior commands.

! Nothing writes `BOT_MODEL` on an installation's behalf any more: an unset `BOT_MODEL` now runs on the middle of the provider's model line rather than a fixed default; `docker-compose.yml` now sets `BOT_MODEL` to empty when it is unset rather than `gpt-5.5`; OpenBot Desktop no longer writes `BOT_MODEL=claude-sonnet-5-5` for an Anthropic key and removes any it previously wrote.

! Team model selection changes: the Let people choose a model switch is removed and replaced by an explicit allowlist under Admin → Providers → Allowed models — teams that previously had the switch on will show no picker until an admin allows models there.

! With Only allowed models on, the new default model (middle of the provider's line) must be explicitly allowed or runs where nobody picked a model are refused.

! A pick on a model no longer allowed or no longer listed by the provider silently falls back to the default rather than running on the chosen model.

! Server and computers must be upgraded together: a 0.3.0 computer has no terminal for the Terminal tab and refuses every hover pointer move the updated live screen now sends.

! The Python LangGraph and Pydantic AI Bots now pin `google-genai` at >=2.34,<3; images built from source after `google-genai` 3.0.0 (2026-11-06) installed 3.x and sent Gemini requests to `/v1` instead of `/v1beta` — rebuilding from this release restores `/v1beta`.

! The Teach a browser workflow is removed along with all routes under `/api/demonstrations`, the instant Skill draft, Save and use with this Bot, and Run on a schedule; recordings made with it remain in the database but are unread.

! Routines can no longer be removed from a Bot or from the server; attempts to remove it or its tools return 403, and an MCP allowlist cannot exclude it.

! With `BOT_PROVIDER=anthropic` and no `BOT_MODEL`, the default model changes from `claude-sonnet-4-5` to `claude-sonnet-5-5`.

! Bot actions that delete, send, share, post, or schedule sends now require approval every time; any Always allow saved for these actions, including those granted on private-share cards, no longer applies.

! Hosted computer browsers now run headed by default (full Chromium on a virtual display), using more memory; set `COMPUTER_BROWSER_MODE=headless` (Helm: `computers.browserMode: headless`) to restore headless mode.

! Bots now share one computer per person; the supervisor stops the old per-Bot containers on start and leaves their volumes for manual removal. Private Bot logins move to the Bot's

owner; people must sign in again to public Bots.

- ! Files on a shared computer previously stored at the workspace root are no longer visible to any Bot; move them into `computers/<person id>~<Bot id>` to restore access.
- ! Single-user OpenBot now refuses requests addressed to unrecognised hostnames and blocks other machines by default; set `OPENBOT_PUBLIC_URL`, `OPENBOT_APP_URL`, or `TRUSTED_ORIGINS` to allow them.
- ! Bot endpoints no longer forward `Authorization` or `x-*` headers from the person's request; remote Bots that read these must be given them as the Bot's own configured headers.
- ! In the `openbot` image, `sudo apt-get` and other package commands are refused; use `sudo openbot-apt install <package>` instead.
- ! With both `COMPOSIO_API_KEY` and `ARCADE_API_KEY` set and no `OPENBOT_BROKER_PROVIDER`, the server refuses to start; set `OPENBOT_BROKER_PROVIDER=composio` or unset the Arcade key.
- ! With both `CPK_INTELLIGENCE_API_KEY` and `INTELLIGENCE_API_KEY` set, the server now uses `CPK_INTELLIGENCE_API_KEY`; remove a stale `CPK_INTELLIGENCE_API_KEY` left over from an older project.
- ! A single-user deployment on Composio re-connects each app once after upgrading: connections recorded under `dev-local-user` are removed from OpenBot (migration `0074_single_user_identity`).
- ! Behind an `OPENAI_BASE_URL` that is not OpenAI or Azure, built-in Bots now use Chat Completions instead of the Responses API; a gateway that serves only the Responses API must set `COPILOTKIT_OPENAI_API=responses`.
- ! Interfaces a Bot drew before the upgrade (A2UI 0.9) no longer render when a conversation is reopened.
- ! An `OPENBOT_PUBLIC_URL`, `OPENBOT_APP_URL`, or `TRUSTED_ORIGINS` value with no scheme (e.g. `example.com:3010`) now stops a single-user server from starting instead of only warning.
- ! A Bot on the Claude Agent SDK harness now starts a new Claude process per turn instead of keeping one per conversation, adding about one to two seconds per turn, and conversations past roughly 100,000 tokens drop their oldest messages.
- ! The built-in safety rules now hand refunds to the person instead of letting the Bot issue them automatically; Bots that issued refunds (e.g. via `SHOPIFY_CREATE_REFUND`) return each one to the user.
- ! Admin → Enterprise controls is removed; its settings are now distributed across Permissions, Providers, Identity providers, Computers, Plugins, Automatic learning, and Audit pages (`/admin/enterprise` redirects to the Admin overview).
- ! OpenBot Desktop for Mac installed from v0.2.0 or earlier cannot self-update; a signed `.dmg` from this release must be installed manually once.
- ! Building the Desktop app from source now requires Rust 1.90 or newer.
- ! OpenBot Desktop no longer connects to a hosted or self-hosted Intelligence; an installation configured that way opens at the first setup screen.
- ! Migration `0070_memory_dedup_per_bot` rebuilds a unique index on `personal_memories` to include the Bot id; while it rolls out, a server still on 0.2.0 cannot save newly formed Bot

memories.

! Migration `0056_sign_in_code_step` requires that a server and its computers be upgraded together; an older computer has no `/sign-in/code` endpoint.

! A Never submit a form boundary rule added before this release still matches only `key == "Enter"`; add the preset again or change the rule to `intent == "activate"` (and update `AGENT_COMPUTER_POLICY` if the line came from `.env.example`).

WAS THIS USEFUL? 👍 👎

CrewAI

Feedback ⓘ 1.15.36 NOTES ↗ RANK

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

> CrewAI expanded its trigger ecosystem to automatically feed external events into crews and flows, added integration tools to call out to Bedrock Agents and other crews, and improved error diagnostics for the `crewai run` CLI.

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

01 Triggers for Gmail, Outlook, Slack and other platforms NEW

55

Crews and flows can now be triggered automatically by payloads from Gmail, Slack, Salesforce, Drive, Outlook, Teams, `OneDrive`, and `HubSpot`. The documentation now includes sample payloads and sample crews in a unified overview covering all supported trigger sources.

– Names all eight platforms but no setup mechanism shown product docs

02 Error class recorded on crewai run load failure IMPROVED

40

`crewai run` now records the error class when it cannot load a project, making load failures easier to diagnose.

– Names exact command and behavior but limited mechanism detail 1.15.36

03 Integration toolkit to call Bedrock Agents and other crews

NEW

38

The updated integration toolkit lets a crew call existing CrewAI automations or Amazon Bedrock Agents directly from within another crew.

– Describes capability but no API or config detail given product docs

WAS THIS USEFUL? 👍 👎

AutoGPT

Feedback ⓘ autogpt-platform-beta-v0.11.1 NOTES ↗

AutoGPT automates multi-step workflows with continuous AI agents that users can build and deploy.

AutoGPT Platform adds `/commands` in copilot chat, input nodes alongside trigger nodes, and background subtasks that return results to chat.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch autogpt-platform-beta-v0.11.1
https://github.com/Significant-Gravitas/AutoGPT.git
# already have the repo? check out this version:
$ git checkout autogpt-platform-beta-v0.11.1
```

Run skills as `/commands` with arguments directly in the copilot chat.

Support input nodes alongside trigger nodes in workflow graphs.

Background subtasks now return results to the chat instead of running silently.

Warn users on canceled trials 3 days and 1 day before access ends.

Generate a per-install Web Push (VAPID) key pair via `make init-env`; set `NEXT_PUBLIC_VAPID_PUBLIC_KEY` to blank to let the frontend use the backend's key automatically.

NEW INTEGRATION Firecrawl Alexandria data tool blocks

NEW INTEGRATION Treg blocks for people, company and web data

UPDATED INTEGRATION Exa integration updated to current API

NEW SUPPORTED PLATFORM Tencent Hy4 added to LLM catalog

└──▷ BREAKING ON UPGRADE

! `.env.default` no longer ships a Web Push (VAPID) key pair and the backend no longer sends push notifications with the old shipped pair. Run `make init-env` to generate a new pair; replacing the pair invalidates all existing push subscriptions. Set `NEXT_PUBLIC_VAPID_PUBLIC_KEY` to blank to have the frontend use the backend's key.

WAS THIS USEFUL? 👍 👎

agentlas-os

Feedback ⓘ os-v1.2.85 NOTES ↗ RANK

Agent OS: keep specialist agents in a hub, spin up a temporary orchestrator per task. Local-first, works with any model.

> agentlas-os improved resilience of workforce runs and agent builds, preserving artifacts and data from failures instead of discarding them, and added clearer diagnostics for blocked browser actions.

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

01 Large JSONL ingestion via source_tail_window

IMPROVED

70

Building an agent from a large JSONL source now uses the `source_tail_window` finding, reading the last 8 MB and last 4000 records instead of refusing oversized inputs, and clearly signals when truncation has occurred.

— Names exact finding and size/record limits. os-v1.2.85

02 Rejected workforce artifacts preserved on verifier failure

IMPROVED

65

A workforce run whose verifier step fails now returns artifacts from completed steps at status `rejected` instead of raising an exception and losing them. The last 2 KB of a failing adapter's stderr is preserved, with secret values masked.

— Explains mechanism and exact stderr retention limit. os-v1.2.85

THINNER COVERAGE BELOW

03 Specific reasons for blocked browser actions

IMPROVED

55

Blocked browser actions now report the specific reason — no desktop app connected, approval timed out, user declined, or answer unreadable — rather than always reporting 'the user did not approve'.

— Lists all four named reasons but no UI/API path given. os-v1.2.85

WAS THIS USEFUL? 👍 👎

LangChain LangGraph

Feedback ⓘ cli==0.4.42 NOTES ↗

RANK

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

> LangGraph's CLI and SDK both gained small operational additions: a metrics command for deployments, and finer control over run resumption and streaming in the SDK.

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

01 **Strict resume and new stream modes for runs** NEW 80

The SDK adds a `strict_resume` parameter to `runs.create`, `runs.stream`, and `runs.wait` to control strict resumption behavior on runs. It also adds `tools` and `lifecycle` values to `StreamMode`, allowing streaming of tool call and lifecycle events.

Resume a paused run and enforce strict resumption semantics so the run fails rather than silently continuing from an inconsistent state.

python

```
result = await client.runs.create(
    thread_id=thread_id,
    assistant_id=assistant_id,
    strict_resume=True,
)
```

Stream tool call events alongside regular output to inspect which tools the agent invokes in real time.

python

```
async for chunk in client.runs.stream(
    thread_id=thread_id,
    assistant_id=assistant_id,
    stream_mode=["tools", "lifecycle"],
):
    print(chunk)
```

— Names exact methods, parameter and stream mode values with runnable code. sdk==0.4.14

02 **Deployment metrics via CLI** NEW 65

The LangGraph CLI adds a `deploy metrics` subcommand to retrieve metrics for deployed agents directly from the command line.

Pull metrics for a running LangGraph deployment to monitor agent activity without leaving the terminal.

```
$ langgraph deploy metrics
```

— Named subcommand with runnable example, but no detail on which metrics. `cli==0.4.42`

WAS THIS USEFUL? 👍 👎

AWS Strands Decider

Feedback ⓘ v0.3.0 NOTES ↗ RANK

A small, fast decision model, or system one model, for agentic workflows. Pick between options or rate on a scale faster than an LLM, with a calibrated confidence on every decision.

> Strands Decider v0.3.0 adds multi-GPU serving, an AWS Lambda deployment option, a torch-free inference engine, and environment-based benchmarks, while removing the ability to disable context-window enforcement.

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

01 **Multi-GPU model splitting at serve time** NEW — 63

Split a model across multiple GPUs at serve time using the `--device-map` flag.

– Names exact flag but no further mechanism detail v0.3.0

THINNER COVERAGE BELOW

02 **Context-window enforcement always on** BREAKING — 55

The `--strict-window` flag is removed; context-window enforcement is now always applied and cannot be disabled.

– Clear before/after and named flag, but no migration steps v0.3.0

03 **Torch-free inference engine** NEW — 50

Run inference without a PyTorch dependency using a new torch-free engine that operates over a backend returning hidden states.

– Explains mechanism but lacks a concrete invocation v0.3.0

04 **AWS Lambda deployment option** NEW — 40

Deploy the model to AWS Lambda with the new Lambda deploy option.

– No specific command or config key given v0.3.0

05 **Environment-based inference benchmarks** NEW — 30

Add environment-based inference benchmarks to measure serving performance.

– Bare description, no command or metric named v0.3.0

└──▷ **BREAKING ON UPGRADE**

! The `--strict-window` flag is removed; context-window enforcement is now always applied and cannot be disabled.

WAS THIS USEFUL? 👍 👎

Nous Research Hermes

Feedback ⓘ v0.23.0 NOTES ↗

CODE ↗ RANK

A terminal-native, self-improving AI agent for autonomous coding and tasks, with persistent memory, agent-created skills, and a multi-platform messaging gateway.

> Hermes v0.23.0 adds a browser-accessible Desktop workspace, simultaneous multi-gateway connections, conversation rewind, paid MCP tool spend tracking, and plugin auto-updates, alongside breaking changes that remove the built-in browser_* tools in favor of browser_exec, rename the MCP server command, raise Desktop's OS requirements, and merge the Office skills.

└──▷ **WHAT SHIPPED · 14 FEATURES** most completely described first ? what's the number?

01 **Removal of built-in browser_* tools** BREAKING 85

The ten built-in `browser_*` tools (`browser_navigate` , `browser_click` , and others) and agent-browser are removed; browser automation now runs through `browser_exec` only. Any prompts, skills, or toolset lists that name the old tools must be updated.

— Names old and new tools with migration note v0.23.0

02 **Browser-accessible Desktop workspace** NEW 83

Enable the full Desktop workspace in a browser by setting `webapp.enabled` in config or toggling Web Access in Settings. It ships in the Docker image and Nix package, off by default.

Expose your Hermes Desktop workspace to a browser on your network or behind a public URL — useful when you want to reach your chats from a device without the Desktop app installed.

> 📍 In Settings, go to Web Access and toggle it on, or add `webapp.enabled: true` to your config.

— Names config key and UI toggle with install surfaces v0.23.0

03 **Configurable subagent delegation strategies** NEW 80

Switch subagent delegation strategies with `/delegation` — keep the default or install the `advanced-delegation` plugin to route each subagent task to a named model or separate profile.

— Names command and plugin with routing mechanism v0.23.0

04 **Conversation rewind** NEW 80

Rewind a conversation to any earlier message in the TUI or classic CLI by pressing Esc twice or typing `/rewind`, then edit and resend or fork a new session from that point.

Roll back to an earlier point in a conversation when a chain of tool calls went the wrong way, then try a different approach without losing the original session.

```
$ /rewind
```

— Runnable command and exact interaction steps v0.23.0

05 **MCP server command renamed** BREAKING 80

`hermes mcp serve` is renamed to `hermes mcp-serve`, moved into a plugin; MCP clients that launch Hermes as a server must update their command.

— Exact old and new commands named v0.23.0

06 **Plugin auto-updates** NEW 78

Update all installed plugins at once with `hermes plugins update`, or enable per-plugin Auto-update toggle in the CLI, chat, Desktop, or dashboard.

— Names exact command and four control surfaces v0.23.0

07 **Curator lock for hand-tuned skills** NEW 75

Lock a hand-curated skill against any agent rewrite, even in YOLO mode, using `hermes curator lock`.

Prevent an agent from overwriting a skill you tuned by hand, even during autonomous YOLO-mode runs.

```
$ hermes curator lock <skill-name>
```

— Runnable command with clear purpose v0.23.0

08 **Office skills merged into one skill** BREAKING 70

The separate Word, PowerPoint, and Excel skills are merged into a single `microsoft-office` skill; old copies remain on disk until manually removed.

— Names new skill and manual cleanup needed v0.23.0

09 **Cron job cancellation and missed-run handling** IMPROVED 65

Cancel a running cron job from the CLI, Desktop, the API, or the agent itself, and configure per-job behavior for runs missed while Hermes was offline.

— Describes behavior but no exact command named v0.23.0

10 **Per-call cost tracking for paid MCP tools** NEW 65

Paid MCP tools can now declare per-call costs; Hermes shows each call's price and the chat's running total, and prompts before a chat exceeds its spend limit, even in YOLO mode.

— Explains mechanism but no config surface named v0.23.0

11 **Inline terminal image rendering** NEW 63

The CLI and TUI now render pasted images and tool screenshots inline — real pixels in kitty, Ghostty, iTerm2, WezTerm, and Konsole, colored blocks elsewhere, and full resolution in VS Code's terminal.

— Names supported terminals but is automatic, no action v0.23.0

12 **Desktop OS requirements raised** BREAKING 63

Desktop now requires macOS 13 or later and 64-bit Windows; macOS 11, macOS 12, and 32-bit Windows are no longer supported after the move to Electron 44.

— Clear version thresholds but no migration step v0.23.0

13 **Simultaneous multi-gateway connections in Desktop** IMPROVED 60

Hermes now keeps all gateways connected simultaneously in one Desktop window, so switching between local and remote (SSH) gateways is instant and Group Chats can mix bots from different machines.

— Clear UI behavior change, no exact config given v0.23.0

THINNER COVERAGE BELOW

14 **New official plugins: Meshy, ElevenLabs, X Chat, REA, bubblewrap**

NEW 55

New official plugins available as one-command installs: Meshy (text and image to 3D), ElevenLabs Music and Sound Effects, X Chat, REA (reverse engineering), and a bubblewrap sandbox.

— Names five plugins but no install command given v0.23.0

└──▷ BREAKING ON UPGRADE

- ! The ten built-in `browser_*` tools (`browser_navigate` , `browser_click` , and others) and agent-browser are removed; browser automation now runs through `browser_exec` only. Any prompts, skills, or toolset lists that name the old tools must be updated.
- ! `hermes mcp serve` is renamed to `hermes mcp-serve` (moved into a plugin); MCP clients that launch Hermes as a server must update their command.
- ! Desktop now requires macOS 13 or later and 64-bit Windows; macOS 11, macOS 12, and 32-bit Windows are no longer supported after the move to Electron 44.
- ! The separate Word, PowerPoint, and Excel skills are merged into a single `microsoft-office` skill; old copies remain on disk until manually removed.

WAS THIS USEFUL? 👍 👎

LangChain Deep Agents

Feedback ⓘ deepagents==0.7.33

NOTES ↗ RANK

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

> LangChain Deep Agents 0.7.33 improves error handling and observability for subagent orchestration, adding default refusal errors, richer async notifications, and task-correlated checkpoints.

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

01 **Checkpoints record task call ID** IMPROVED 55

Subagent checkpoints now record the `task` call ID, making it possible to correlate a checkpoint with the specific `task` invocation that produced it.

— Names the exact field added and its correlation use deepagents==0.7.33

02 **Model refusal errors surfaced to parent agent** IMPROVED 49

Deep Agents now raises errors on model refusals by default and reports subagent refusals up to the parent agent, so orchestration code can react to blocked calls rather than silently continuing.

— Describes new default behavior but no config flag to control it deepagents==0.7.33

03 **Notification metadata on async subagent messages** IMPROVED 28

Async subagent notifications now carry notification metadata, giving consumers richer context on each notification event.

— Vague on what metadata fields were actually added deepagents==0.7.33

WAS THIS USEFUL? 👍 👎

HeyGen HyperFrames

Feedback ⓘ v0.9.17 NOTES ↗ RANK

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

> HyperFrames tightened CLI render and inspection tooling with disk-saving windowed frame extraction, budget-bounded checks, and clip auditing, while also adding reference-background reporting and smarter marquee selection in Studio.

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

01 **Time-bounded checks with check --budget** NEW 77

The `check --budget` flag lets CI fail fast by running checks only as long as the given budget allows, then lets you inspect partial results, e.g. `npx hyperframes check --budget 30s`.

Fail CI fast by running checks only as long as your budget allows, then inspecting partial results.

```
$ npx hyperframes check --budget 30s
```

— Runnable command with exact flag and behaviour described. v0.9.17

02 **Clip listing via info --media** NEW 77

`info --media` lists every audio/video clip a render uses, with timestamps on the film's clock, useful for auditing a render before submitting to cloud render, e.g. `npx hyperframes info --media my-film.hf`.

Audit every clip a render hears or sees, with timestamps on the film's clock, before submitting to cloud render.

```
$ npx hyperframes info --media my-film.hf
```

— Runnable command with exact subcommand and clear purpose. v0.9.17

03 **Windowed frame extraction via HF_VIDEO_FRAME_WINDOW** NEW 68

Setting the `HF_VIDEO_FRAME_WINDOW` environment variable to `true` reduces disk usage during long renders by making each capture segment extract only the frames it needs and delete them afterward.

— Names exact env var and mechanism, but no example command given. v0.9.17

THINNER COVERAGE BELOW

04 **reference_background exposed in CLI render output** NEW

56

The CLI now prints a render's `reference_background` as an `@hf-result` line and includes it in batch manifest rows.

— Names exact field and output format but no usage example. v0.9.16

05 **Alt-marquee deselection in Studio** IMPROVED

37

Studio now removes covered items from the current selection when using Alt marquee.

— Behaviour named but no exact UI path or mechanism detail. v0.9.16

WAS THIS USEFUL? 👍 👎

LiveKit Agents

Feedback ⓘ livekit-agents@1.8.11 NOTES ↗

RANK

LiveKit Agents builds realtime voice and video AI agents on LiveKit's WebRTC infrastructure, with turn detection, interruption handling, and telephony.

> LiveKit Agents adds a Responses API LLM backed by LiveKit Inference with WebSocket connection reuse and configurable reasoning effort, alongside a wave of new STT/TTS/translation provider integrations (Zoom AI, Kalpa, Rime, Fish Audio, OpenAI, Google, Respeecher) and small config-option additions to the existing Speechmatics, Rime, and CambAI integrations.

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

01 **Responses API via LiveKit Inference** NEW 93

Adds `inference.responses.LLM` to use the Responses API through LiveKit Inference credentials, supporting structured output and web search; pass `use_websocket=True` to reuse the connection across conversation turns instead of reconnecting each time. A new `reasoning=` parameter on both `inference.responses.LLM` and `inference.LLM` controls reasoning effort, and passing `reasoning=None` keeps the provider's own default rather than the new minimum-effort default that now applies otherwise.

Use the Responses API backed by LiveKit Inference with WebSocket transport to reuse the connection across conversation turns.

python

```python
from livekit.agents import inference

llm = inference.responses.LLM("openai/gpt-4.1",
use_websocket=True)
```

Pin reasoning effort to a specific level, or pass `reasoning=None` to let the provider choose its own default instead of the per-model minimum.

python

```python
from livekit.agents import inference

# Use lowest effort (new default)
llm = inference.responses.LLM("openai/gpt-4.1")

# Delegate effort choice to the provider
llm = inference.responses.LLM("openai/gpt-4.1", reasoning=None)
```

— Runnable code shows the class, flag, and parameter defaults. `livekit-agents@1.8.11`

02 New STT, TTS, and translation provider integrations NEW

70

Adds a Zoom AI STT plugin with live streaming and per-utterance recognition; a Kalpa TTS plugin supporting the `kalpa-tts-1` streaming model in English and Hindi; Rime STT and Rime Prism speech-to-speech through the standard realtime interface (requires Python 3.11+; STT requires the `stt` extra); and Fish Audio realtime STT using the `transcribe-stream-1` model with word timings and recognition context. OpenAI adds `gpt-realtime-translate` through a beta translation interface with `gpt-realtime-whisper` as input transcription, and Google adds `gemini-3.5-live-translate-preview` translation through the Gemini API and Vertex AI. Respeecher now exposes public Swedish, Brazilian Portuguese, German, and Arabic TTS models.

— Names each new plugin and model but gives no usage example. `livekit-agents@1.8.11`

THINNER COVERAGE BELOW

03 Provider option additions for Speechmatics, Rime, and CambAI

IMPROVED

55

Speechmatics STT now supports 8 kHz audio and an `emit_sentences` config option. Rime adds a `custom_lexicon` option on its Coda integration. CambAI TTS now accepts a `speaking_rate` option.

— Names exact config options but gives no usage example. `livekit-agents@1.8.11`

WAS THIS USEFUL? 👍 👎

Graphicle AI Graphicle

Feedback ⓘ v0.1.13 NOTES ↗

RANK

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

> Graphicle v0.1.13 introduces a project wiki subsystem that auto-generates and maintains per-area documentation from the code graph, alongside early-stop graph update optimizations and a required one-time migration after upgrading.

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

01 Model-generated project wiki with build, update, hook commands

NEW

86

`graphicle wiki build` generates a model-written wiki with one article per code area, stored in `.graphicle/wiki/written/`; spend is capped with `--max-cost` (default 1 USD), previewable with `--dry-run`, and trackable via `graphicle usage`. `graphicle wiki update` keeps it current by rewriting only articles whose code, docs, or reviewed memories changed, and `graphicle wiki hook on` adds an optional post-commit hook to automate updates; a `graphicle wiki check` command is also referenced.

— Names exact commands, flags, default cost, and storage path v0.1.13

02 Required migration after upgrading to v0.1.13

BREAKING

75

Because the extraction cache schema changed, the first `graphicle-graph update` after upgrading performs a full update rather than an incremental one since the early-stop record does not yet exist. Users must also run `graphicle sync` once per project after upgrading, as the saved graph and query file are keyed to the new build and must be rewritten.

— Gives exact commands and reason for required migration step v0.1.13

THINNER COVERAGE BELOW

03 Early-stop graph updates IMPROVED

25

Graphicle adds early-stop graph updates, an optimization to the code graph update process mentioned alongside the wiki release.

— Mentioned only by name, no mechanism given v0.1.13

└──▷ **BREAKING ON UPGRADE**

! The first `graphicle-graph update` after upgrading performs a full update because the extraction cache schema changed and the early-stop record does not exist yet.

! Run `graphicle sync` once in each project after upgrading, as the saved graph and query file are keyed to the new build and must be rewritten.

WAS THIS USEFUL? 👍 👎

Feynman

Feedback ⓘ v0.5.64 NOTES ↗ RANK

Feynman is a personal AI research assistant that searches papers and the web, builds literature reviews, and produces cited research reports.

> Feynman expands its search backend with two new providers, raises GitHub README fetch limits, brings parallel-researcher execution to `/lit`, and warns that pre-upgrade background runs can no longer be resumed.

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

01 **New sofya and outcrawl search providers** NEW 75

Feynman adds `sofya` and `outcrawl` as selectable search providers via `feynman search set`; they are used only when explicitly chosen.

Switch to the new `sofya` search provider for a research session that benefits from its index.

```
$ feynman search set sofya
```

— Names exact command and providers but not how they differ. v0.5.64

THINNER COVERAGE BELOW

02 **GitHub README fetch limit raised to 100K characters** IMPROVED 55

GitHub repository fetching now returns READMEs up to 100K characters instead of the previous 8K limit.

— Clear before/after numbers but no user action required. v0.5.64

03 **/lit command runs parallel researchers** IMPROVED 50

The `/lit` command now launches parallel researchers using the same workflow as `/deepresearch`.

— Names both commands but no further mechanism detail. v0.5.64

04 **Background runs require relaunch after upgrade** BREAKING 45

Background runs started before v0.5.64 must be relaunched — resuming them is no longer supported, due to a trust change in files those runs could rewrite.

— States migration requirement but omits trust mechanism detail. v0.5.64

└──▷ **BREAKING ON UPGRADE**

! Background runs started before v0.5.64 must be relaunched — resuming them is no longer supported due to a trust change in files those runs could rewrite.

WAS THIS USEFUL?  

truefoundry TrueForge

Feedback ⓘ

@truefoundry/trueforge-ui@0.184.1 NOTES ↗ RANK

The open-source agent harness - the runtime layer that turns an LLM into a working agent.

> TrueForge's headline shipment this window is a full long-term memory system backed by Zep, with a new tenant-wide web search toggle, sandbox-environment capability gating, and a handful of smaller API, client and UI refinements.

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

01 **Long-term memory for agents via Zep** NEW 98

A new Memory toggle in the agent builder (off by default) gates long-term memory across chats, ingestion, and agent specs. Configure a provider via Settings → Memory or the `/api/v1/settings/memory-providers` API in self-hosted OSS deployments, or via `TRUEFOUNDRY_MEMORY_ENABLED`, `TRUEFOUNDRY_MEMORY_PROVIDER` (supports `zep`), `TRUEFOUNDRY_MEMORY_API_KEY`, optional `TRUEFOUNDRY_MEMORY_SERVER_URL`, optional `TRUEFOUNDRY_MEMORY_SETTINGS` JSON, and `TRUEFOUNDRY_MEMORY_TENANT_ALLOWLIST` in TrueFoundry-managed mode. Agents with `config.memory.enabled` receive `memory_search` and `memory_add` tools, with the user's memory summary injected as a hidden message capped at ~1,000 tokens on the first turn; a new `memory_ingest_cursor` table tracks ingestion progress and a `memory-ingest` controller loop sends finished UI chats and sessions from registry agents with memory enabled to Zep, forwarding user messages, `ask_user_question` answers, and the final agent reply.

Turn on memory for a specific agent once a provider is configured — navigate to the agent builder and flip the Memory toggle.

📍 1. Open the TrueForge UI and go to the agent builder for your agent. 2. Under agent settings, locate the Memory toggle (visible only when a memory provider is active). 3. Switch it on to enable memory for chats, ingestion, and the agent spec.

Register Zep as the tenant memory provider in a self-hosted OSS deployment so agents can persist and recall conversation memory.

```
$ POST /api/v1/settings/memory-providers
```

Supply Zep credentials to a TrueFoundry-managed instance at startup without touching the UI.

```text
TRUEFOUNDRY_MEMORY_* (set the relevant env vars in your
deployment environment before starting TrueForge)
```

Turn on memory for the whole tenant so all chats, ingestion pipelines, and agent specs can use it.

> 📍 In the TrueForge UI, go to Settings → Memory and enable the tenant-wide Memory toggle (off by default). The toggle appears only after a memory provider is configured.

Turn on the Memory toggle for an agent once the provider is registered, to enable memory across chats and ingestion.

> 📍 In the agent builder, go to Settings → Memory and enable the tenant-wide memory switch, then open your agent spec and toggle Memory on.

Turn on tenant-wide memory once the Zep provider is configured, to enable memory in chats, ingestion, and agent specs.

> 📍 In the TrueForge admin UI, go to Settings › Memory and enable the tenant-wide Memory toggle.

Turn on long-term Zep memory for all tenants in a TrueFoundry-mode deployment.

```ini
TRUEFOUNDRY_MEMORY_ENABLED=true
TRUEFOUNDRY_MEMORY_PROVIDER=zep
TRUEFOUNDRY_MEMORY_API_KEY=<your-zep-api-key>
TRUEFOUNDRY_MEMORY_SERVER_URL=https://zep.example.com
TRUEFOUNDRY_MEMORY_TENANT_ALLOWLIST=
```

Enable memory on an agent so its chat sessions receive `memory_search` and `memory_add` tools and persist conversation history to Zep.

```yaml
config:
  memory:
    enabled: true
```

Activate Zep-backed long-term memory for all tenants in TrueFoundry mode by setting the required environment variables before starting TrueForge.

```text
TRUEFOUNDRY_MEMORY_ENABLED=true
TRUEFOUNDRY_MEMORY_PROVIDER=zep
TRUEFOUNDRY_MEMORY_API_KEY=<your-zep-api-key>
TRUEFOUNDRY_MEMORY_SERVER_URL=https://zep.example.com
TRUEFOUNDRY_MEMORY_TENANT_ALLOWLIST=
```

Restrict memory ingestion to specific tenants and pass custom Zep loop settings via `TRUEFOUNDRY_MEMORY_SETTINGS`.

```text
TRUEFOUNDRY_MEMORY_ENABLED=true
TRUEFOUNDRY_MEMORY_PROVIDER=zep
TRUEFOUNDRY_MEMORY_API_KEY=<your-zep-api-key>
TRUEFOUNDRY_MEMORY_TENANT_ALLOWLIST=tenant-a,tenant-b
TRUEFOUNDRY_MEMORY_SETTINGS={"loop_interval_seconds": 60}
```

Turn on memory for a specific agent so its sessions get `memory_search` and `memory_add` tools automatically.

```yaml
config:
  memory:
    enabled: true
```

— Names every env var, endpoint, config key and tool `@truefoundry/trueforge-ui@0.184.1` `@truefoundry/trueforge@0.184.1` `@truefoundry/trueforge-sdk@0.184.1` `@truefoundry/trueforge-assistant-ui-run…` `@truefoundry/trueforge@0.184.0` `@truefoundry/trueforge-core@0.184.0`

02

Capability gate for sandbox environments NEW

70

A new `sandbox.environments.enabled` flag in Capabilities gates the Environments UI and CRUD operations; when the shared sandbox has no snapshot or environment concept, the UI hides Environments and CRUD calls return HTTP `422`.

Check which settings sections are available and whether web search is currently enabled for the tenant.

```
$ curl -X GET https://<host>/api/v1/settings/configuration \
  -H 'Authorization: Bearer <admin-token>'
```

— Names the flag and the error code, no UI path `@truefoundry/trueforge-ui@0.184.0`

`@truefoundry/trueforge@0.184.0`

03 Sandbox provider type stored on environments IMPROVED

65

Sandbox environments created or updated in TrueFoundry mode now store `type: truefoundry` alongside the backing `sandbox_provider` (e.g. `daytona`), with a migration that backfills `type` on existing environment versions.

— Describes schema change and migration but no reader action

`@truefoundry/trueforge@0.184.0`

THINNER COVERAGE BELOW

04 User-agent header on outbound requests NEW

55

TrueForge now prepends `user-agent: trueforge/<version>` on outbound model, MCP, and ServiceFoundry fetches via `OutboundFetchOptions.userAgent`, letting TrueFoundry identify TrueForge traffic and its version.

— Names the option and header but no usage example `@truefoundry/trueforge@0.184.0`

05 Default model HTTP timeout settings IMPROVED

55

Core client and server configuration now sets a default model HTTP header timeout of 20 seconds, keeping the body timeout at 120 seconds, with the environment override preserved.

— Exact timeout values given but no config key named `@truefoundry/trueforge-core@0.184.0`

06 Token breakdown in session tooltips IMPROVED

40

Session token tooltips now lead with total tokens, then break out input, output, cache read, and cache write as separate rows with exact counts, with one bar shown for the input breakdown only.

— Describes UI change but no navigation path given `@truefoundry/trueforge-ui@0.184.0`

07 Exported ask-user-question tool name constant NEW

35

Core now exports `ASK_USER_QUESTION_TOOL_NAME` from core, making the tool name constant available to consumers.

— Bare export name with no usage context `@truefoundry/trueforge@0.184.0`

└──▷ BREAKING ON UPGRADE

! Catalog settings routes that the deployment does not own now answer 422 naming the route, where they previously answered 403.

! Catalog settings routes that the deployment does not own now return 422 naming the route, where they previously returned 403.

! Catalog settings routes that the deployment does not own now return 422 (naming the route) instead of 403 — any client code that branches on 403 for unowned catalog settings routes will need updating.

WAS THIS USEFUL? 👍 👎

Microsoft Semantic Kernel

Feedback ⓘ python-1.47.0

NOTES ↗ RANK

Semantic Kernel provides an open-source SDK for building AI agents and integrating language models into C#, Python, and Java applications.

> Semantic Kernel's Python SDK added NVIDIA nvext passthrough, Gemini thinking-token usage counting, automatic Google AI embedding batching, and improved JSON schema generation, while dropping Python 3.10 support.

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

01 **NVIDIA nvext settings passthrough** NEW 65

NVIDIA-backed chat completion requests can now send NVIDIA `nvext` settings via `extra_body`, allowing provider-specific options to reach the underlying API.

– Names exact field and parameter but no example python-1.47.0

THINNER COVERAGE BELOW

02 **Google AI embedding request batching** IMPROVED 55

Google AI embedding requests now batch automatically to stay within the provider's 100-input limit, removing the need for manual chunking of embedding calls.

– Names concrete limit but no code sample python-1.47.0

03 **Python 3.10 support removed** BREAKING 55

Python 3.10 is no longer supported; the minimum required Python version is now 3.11, with CI now covering Python 3.11 through 3.14.

– Clear breaking change with version numbers python-1.47.0

04 **storage_name for vector store hybrid search** IMPROVED 50

Vector store collections can now use `storage_name` for hybrid search fields and vector queries.

– Names config field but lacks usage example python-1.47.0

05 **Gemini thinking-token usage counting** IMPROVED 45

Gemini thinking tokens now count toward usage totals in Google AI completion responses, giving more accurate token accounting for reasoning models.

– Explains behaviour change but no config or API detail python-1.47.0

06 Array schema for variadic tuple parameters IMPROVED

35

Variadic tuple parameters now produce an array schema in generated JSON schemas, improving function-calling schema accuracy for such parameters.

— Behaviour named but no usage path given `python-1.47.0`

└──▷ **BREAKING ON UPGRADE**

! Python 3.10 is no longer supported; the minimum required Python version is now 3.11.

WAS THIS USEFUL? 👍 👎

OpenClaw

Feedback ⓘ v2026.11.1 NOTES ↗ CODE ↗ RANK

OpenClaw is a personal AI assistant platform that connects messaging channels, tools, and model providers through a gateway.

> OpenClaw's single release this window adds opt-in native Agent and Workflow orchestration for Claude CLI, optional Anthropic workspace ID recording, smarter Bedrock Mantle model routing, and a cluster of reliability fixes to Google Interactions requests.

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

01 Native Agent and Workflow orchestration in Claude CLI NEW

88

Operators can opt into native Agent and Workflow orchestration by setting `plugins.entries.anthropic.config.managedNativeTools`; the default remains off, and restricted runs keep their existing tool limits.

Enable native Agent and Workflow orchestration for a Claude CLI operator deployment that needs managed tool access beyond the default restricted set.

yaml

```yaml
plugins:
  entries:
    anthropic:
      config:
        managedNativeTools: true
```

— Named config key, runnable YAML example, default behavior stated v2026.11.1

02 Google Interactions request reliability and compatibility fixes

IMPROVED

65

Google Interactions requests now honor managed proxy and TLS settings, rotate eligible API keys before streaming after hitting a rate limit, preserve native video input, and forward nucleus sampling consistently; use `google-generative-ai` when explicit Gemini cache handles are required.

— Several named behaviors and one named provider option, no config keys v2026.11.1

THINNER COVERAGE BELOW

03 Bedrock Mantle model-route selection IMPROVED

57

Bedrock Mantle now automatically routes discovered numbered OpenAI GPT models to the Responses route, while GPT-OSS models keep Chat Completions and Claude keeps Anthropic Messages.

— Clear mechanism and scope but no config flag to act on v2026.11.1

04 **Optional Anthropic workspace ID for API-key setup** NEW

48

Anthropic API-key setup now accepts an optional Anthropic workspace ID; existing workspace headers are preserved and workspace-scoped keys require no extra setting.

— Explains behavior but no concrete field name or UI path given v2026.11.1

05 **Active context budget shown in Control UI model picker**

IMPROVED

40

The Control UI model picker now shows the configured active context budget rather than only the model maximum.

— Names exact UI location but no further mechanism v2026.11.1

06 **Saved subscription accounts in Claude CLI** IMPROVED

35

Claude CLI now allows saved subscription accounts to make configured models available without requiring a separate native login.

— States behavior change but no setup detail given v2026.11.1

07 **Visible links for Google Chat remote attachments** IMPROVED

32

Google Chat now delivers remote HTTP(S) attachments as visible links alongside captions.

— Describes behavior but nothing a reader configures v2026.11.1

WAS THIS USEFUL? 👍 👎

octop-harness

Feedback ⓘ v1.0.8 NOTES ↗

A production-grade agent runtime engineered from the Harness Engineering philosophy.

File tools now accept workspace-relative paths: `base_dir` resolves relative to the document directory; `execute`'s `cwd` scopes to a single command.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v1.0.8 https://github.com/TencentCloud/octop-harness.git
# already have the repo? check out this version:
$ git checkout v1.0.8
```

NEW INTEGRATION Tencent Cloud TokenHub provider preset with HY4 model

WAS THIS USEFUL? 👍 👎

◆ MCP Servers & Tooling

seo-monster

Feedback ⓘ v0.10.0 NOTES ↗ RANK

SEOMonster: a local, user-credential-driven MCP server for SEO workflows over Google Search Console, GA4, PageSpeed Insights, and Cloudflare. Read-first, open source.

> seo-monster's v0.10.0 release adds Unicode-aware text analysis and GA4 filter validation, hardens URL fetching against untrusted sources, and raises its minimum SDK and Python version requirements.

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

01 **Updated SDK and Python version requirements** BREAKING

65

MCP Python SDK v1 is no longer supported; SDK 2.3 or later (below v3) is now required. Python 3.11 or later is also now required, and earlier Python versions will not work.

– Exact version floors and ceiling given, clear upgrade action v0.10.0

THINNER COVERAGE BELOW

02 **GA4 filter validation before submission** NEW 40

seo-monster now validates GA4 filters before submission, catching invalid filters prior to sending requests.

– Names GA4 surface but no validation rules or errors shown v0.10.0

03 **Safer provider-selected URL fetching** IMPROVED 35

seo-monster now applies safer provider-selected URL fetching to reduce exposure from untrusted sources.

– States security goal but no mechanism or affected endpoints named v0.10.0

04 **Unicode lexical extraction for text analysis** NEW 30

seo-monster now supports Unicode lexical extraction for text analysis, improving handling of non-ASCII content.

– Named capability only, no mechanism or scope given v0.10.0

└──▷ BREAKING ON UPGRADE

! MCP Python SDK v1 is no longer supported; SDK 2.3 or later (below v3) is required.

! Python 3.11 or later is now required; earlier Python versions will not work.

WAS THIS USEFUL? 👍 👎

gflow-cli

Feedback ⓘ v0.93.0 NOTES ↗ RANK

Drive Google Flow from the command line: Veo video and Imagen images, scripted, batched and pipeline-ready. Ships an MCP server so coding agents can drive it too, giving you and your agents full programmatic control over generation.

> gflow-cli adds a short alias for the Nano Banana 2.1 image model, simplifying image generation commands.

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

01 **`nano21` alias for Nano Banana 2.1 model** NEW 65

Image generation commands can now reference the Nano Banana 2.1 model using the shorter alias `nano21` instead of its full name, e.g. `gflow image t2i "a sunlit coastal village at golden hour" --model nano21`.

Generate an image using the Nano Banana 2.1 model via its new short alias instead of the full name.

```
$ gflow image t2i "a sunlit coastal village at golden hour" --model nano21
```

— Exact flag and alias named with runnable command, but minimal mechanism. v0.93.0

WAS THIS USEFUL? 👍 👎

concord-mcp

Feedback ⓘ v0.20.0 NOTES ↗ RANK

Live messaging for coding agents

> concord-mcp added Conductor support, enabling idle wakeup for SDK-hosted Claude Code agents.

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

01 **Conductor support for idle wakeup** NEW 35

concord-mcp now supports Conductor, enabling idle wakeup for SDK-hosted Claude Code agents; the Codex host is not supported in Conductor.

— Names agent/host scope but no setup steps or config shown. v0.20.0

WAS THIS USEFUL? 👍 👎

api-mcp-server

Feedback ⓘ v3.30.0 NOTES ↗ RANK

MCP server that exposes Hostinger's platform API for hosting deployment, CDN control, domain management, and WordPress operations.

> api-mcp-server consolidates PHP management in its hosting module with a new single-call update operation while streamlining the older granular operations.

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

01 **Consolidated PHP management in hosting module** IMPROVED 88

The new `hosting_php_update` operation updates a website's PHP version, extensions, and options in one call. The four previous PHP operations — `hosting_php_reset-extensions`, `hosting_php_update-extensions`, `hosting_php_update-options`, and `hosting_php_update-version` — are now hidden from search but remain callable via `execute`.

— Names exact operations and new consolidated call, no further mechanism given v3.30.0

WAS THIS USEFUL? 👍 👎

Composio

Feedback ⓘ announced · Nov 3, 2026 BLOG ↗ RANK

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

> Composio added Tavily to its Instant Toolkits, letting connected agents run web search and content extraction without a separate Tavily account or API key.

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

01 **Tavily integration in Instant Toolkits** NEW 77

Composio's Instant Toolkits now support Tavily, exposing the Search and Extract actions so agents can find web results (titles, URLs, content) and retrieve full page content from URLs. No separate Tavily account, API key, or subscription is needed — calls are billed per use from the organisation's Composio wallet, and session policies control which eligible tools the agent can use. Enable it via project settings by turning on the Tavily toolkit under Instant Toolkits.

Enable Tavily in an existing Composio project so agents can run web searches and page extractions without a separate Tavily account.

> 📍 In your project settings, go to Instant Toolkits and enable the Tavily toolkit. Confirm your organisation has a prepaid Composio wallet balance. Session policies control which eligible tools the agent can use.

— Names exact actions and setup steps but UI path not a command `launch-20261103-538f4827`

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Agentic Infrastructure

Perplexity API

Feedback ⓘ updated · Nov 6, 2026 NOTES ↗

RANK

Perplexity API provides language models with real-time web search and retrieval for building applications that research and answer questions.

> Perplexity's Agent API gains a new `ultra` effort level for broad research and data-gathering tasks, running on claude-opus-5-5 with background mode support for long-running jobs.

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

01 **`ultra` effort level in Agent API** NEW 86

Pass `effort="ultra"` (or the pinned version `ultra@2026-11-05.0`) to the Agent API to run broad research and data-gathering tasks, with support for background mode on long-running runs. The `ultra` effort level runs on `anthropic/claude-opus-5-5` with no frozen configuration, and can be called by name or pinned to the returned version to keep behaviour stable across future model changes.

Pin to a specific `ultra` snapshot so future model changes do not alter your run behaviour.

```
$ curl -X POST https://api.perplexity.ai/agents \
  -H 'Authorization: Bearer <token>' \
  -H 'Content-Type: application/json' \
  -d '{"effort": "ultra@2026-11-05.0", "messages": [{"role":
"user", "content": "Map all known threat actor TTPs targeting
cloud storage in 2026."}]}'
```

— Names exact parameter, pinned version format, backing model, and runnable example.

snapshot-20261106

WAS THIS USEFUL? 👍 👎

Google Gen AI JavaScript SDK

Feedback ⓘ v3.0.0

NOTES ↗ RANK

The Google Gen AI JavaScript SDK provides TypeScript and JavaScript interfaces to Gemini APIs and live model sessions.

> Google Gen AI JavaScript SDK v3.0.0 ships explicit API-version control, widens code-execution results to multimodal content, and raises its minimum Node.js requirement to 22 — with the Gemini Developer API now defaulting to v1 instead of v1beta.

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

01 Explicit API version control via env var and client option

90

BREAKING

Set the `GOOGLE_GENAI_API_VERSION` environment variable (accepting `beta`, `v1`, `v1beta`, or `v1beta1`) to control the API version for both Gemini Developer and Vertex AI clients; `apiVersion` or `httpOptions.apiVersion` passed to `GoogleGenAI` in code takes precedence when set. This pairs with a default change: Gemini Developer API clients now default to `v1` instead of `v1beta`, so callers who need the old behavior must set `GOOGLE_GENAI_API_VERSION=beta` or pass `apiVersion: 'v1beta'`; Vertex AI clients (`vertexai: true`) are unaffected and keep defaulting to `v1beta1`.

Pin the Gemini Developer API back to `v1beta` after the new `v1` default, without changing any code.

```
$ export GOOGLE_GENAI_API_VERSION=v1beta
```

Override the API version in code so it takes precedence over the environment variable.

typescript

```typescript
import { GoogleGenAI } from '@google/genai';

const ai = new GoogleGenAI({
  apiKey: process.env.GEMINI_API_KEY,
  apiVersion: 'v1beta',
});

const response = await ai.models.generateContent({
  model: 'gemini-flash-latest',
  contents: 'Summarize this document.',
});
console.log(response.text);
```

— Names env var, flags, and exact migration command `v3.0.0`

02 Multimodal content in code-execution results

BREAKING

75

`CodeExecutionResultStep.result` is now typed as `Array<Content> | string` via the new `CodeExecutionResultStepResult` union, so code-execution steps can return rich content (e.g. images, charts) alongside plain text instead of only strings. TypeScript code reading `step.result` must narrow the union with `typeof step.result === 'string'` to continue compiling.

Handle multimodal code-execution results after the `CodeExecutionResultStep.result` type widened to `Array<Content> | string`.

typescript

```typescript
import { Content } from '@google/genai';

for (const step of codeExecutionSteps) {
  if (typeof step.result === 'string') {
    console.log('text result:', step.result);
  } else {
    // Array<Content> — may contain images, charts, etc.
    for (const part of step.result as Content[]) {
      console.log('multimodal part:', part);
    }
  }
}
```

— Type name and narrowing pattern given, no broader context `v3.0.0`

THINNER COVERAGE BELOW

03 **Minimum Node.js version raised to 22** BREAKING — 55

The SDK now requires Node.js 22 as a minimum (was 20). Users who must stay on Node.js 20 should pin to `@google/genai@^2`.

— Clear constraint and pin workaround given v3.0.0

04 **Audio transcription config additions** NEW — 45

Adds `source_text` to `AudioTranscriptionConfig` and a new `SmartTranscriptionMode` to configure audio transcription behavior.

— Names fields but no mechanism or usage shown v3.0.0

05 **CLI transcribe default model updated** IMPROVED — 35

The CLI `transcribe` command now defaults to `gemini-3.5-transcribe` as its model.

— Names command and new default, no further detail v3.0.0

06 **MALFORMED_RESPONSE finish reason** NEW — 30

Adds `MALFORMED_RESPONSE` to the `FinishReason` enum to surface malformed-response terminations explicitly.

— Bare enum addition, no usage context v3.0.0

07 **Office and OpenDocument MIME type input support** NEW — 27

Interactions now accept Office and OpenDocument MIME types as input content.

— No specific MIME type list or example given v3.0.0

└──▷ BREAKING ON UPGRADE

! Gemini Developer API clients now default to API version `v1` instead of `v1beta`. Restore the old default with `export GOOGLE_GENAI_API_VERSION=beta` or pass `apiVersion: 'v1beta'` to `GoogleGenAI`. Vertex AI clients (`vertexai: true`) are unaffected and continue to default to `v1beta1`.

! Minimum supported Node.js version is now 22 (was 20). Users who must stay on Node.js 20 should pin to `@google/genai@^2`.

! `CodeExecutionResultStep.result` is now typed as `Array<Content> | string` (`CodeExecutionResultStepResult`) instead of `string`. TypeScript code that reads `step.result` as a plain string must narrow the union with `typeof step.result === 'string'`.

WAS THIS USEFUL? 👍 👎

Google Gen AI Python SDK

Feedback ⓘ v3.0.0 NOTES ↗

RANK

The Google Gen AI Python SDK provides Python interfaces to Gemini model generation, streaming, tools, and embeddings.

> Google Gen AI Python SDK v3.0.0 overhauls HTTP transport defaults (switching to `httpx2`) and API version control via a new environment variable, alongside smaller additions for audio transcription, malformed-response detection, Office file support, and multimodal code-execution results — several of which are breaking changes.

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

01 HTTP transport switch to httpx2

BREAKING

100

The SDK's default HTTP transport is now `httpx2`, which uses the OS certificate trust store (via `truststore`) instead of `certifi` when `SSL_CERT_FILE` / `SSL_CERT_DIR` are not set; `httpx` is no longer installed by default. Restore the legacy transport by installing `google-genai[httpx]` and setting `GOOGLE_GENAI_HTTP_CLIENT=httpx`. New `httpx_client` and `httpx_async_client` fields on `HttpOptions` let you pass `httpx.Client` or `httpx.AsyncClient` instances directly (also requires `google-genai[httpx]`). Package-specific `httpx` objects passed via HttpOptions(client_args=..., async_client_args=...) — such as `Timeout`, `Limits`, `MockTransport`, or custom `Auth` subclasses — and transport exceptions caught from `client.models`, `client.files`, etc. must now match the active HTTP client.

Restore the legacy `httpx` transport for environments that rely on `httpx`-specific transports or auth classes.

```
$ pip install "google-genai[httpx]"
export GOOGLE_GENAI_HTTP_CLIENT=httpx
```

— Names env vars, install extra, fields, and affected objects v3.0.0

02 Environment-variable control of Gemini API version

BREAKING

95

The new `GOOGLE_GENAI_API_VERSION` environment variable lets you set the Gemini API version (`beta`, `v1`, or `v1beta`) across all clients without code changes, e.g. `export GOOGLE_GENAI_API_VERSION=v1`. The default Gemini Developer API version changed from `v1beta` to `v1`; restore the old default with

`GOOGLE_GENAI_API_VERSION=v1beta` or by passing types.HttpOptions(api_version='v1beta') to `genai.Client`.

Pin all API calls to the stable `v1` endpoint in CI without touching application code.

```
$ export GOOGLE_GENAI_API_VERSION=v1
```

— Names env var, values, and exact migration code `v3.0.0`

---

03 **Multimodal code execution results** BREAKING 75

`CodeExecutionResultStep.result` changed from a plain `string` to the new `CodeExecutionResultStepResult` union type, enabling multimodal content instead of just strings. Code that reads `result` as a plain string must now check which variant is present and extract the string from it.

— Explains before/after and migration need, names the type `v3.0.0`

THINNER COVERAGE BELOW

04 **Smart audio transcription mode** NEW 45

`AudioTranscriptionConfig` gains a `source_text` field, and audio transcription can now use the new `SmartTranscriptionMode`.

— Names fields but no usage or behaviour detail `v3.0.0`

05 **Malformed response detection in FinishReason** NEW 40

`FinishReason` adds a new `MALFORMED_RESPONSE` value so callers can detect malformed model responses.

— Names the enum value but no further mechanism `v3.0.0`

06 **Office and OpenDocument MIME support in interactions** NEW 40

New MIME types let you send and receive Office and OpenDocument file formats via `interactions`.

— Names the surface but no specific MIME types listed `v3.0.0`

└──▷ **BREAKING ON UPGRADE**

! The default Gemini Developer API version changed from `v1beta` to `v1`; restore the old default by setting `GOOGLE_GENAI_API_VERSION=v1beta` or passing types.HttpOptions(api_version='v1beta') to `genai.Client`.

! `httpx` is no longer installed by default; `httpx2` is now the required dependency and default HTTP client. Install `google-genai[httpx]` and set `GOOGLE_GENAI_HTTP_CLIENT=httpx` to restore the previous behaviour.

! `httpx2` uses the OS certificate trust store (via `truststore`) instead of `certifi` by default when `SSL_CERT_FILE` and `SSL_CERT_DIR` are not set.

! Package-specific `httpx` objects passed via HttpOptions(client_args=..., async_client_args=...) — such as `Timeout`, `Limits`, `MockTransport`, or custom `Auth` subclasses — and transport exceptions caught from `client.models`, `client.files`, etc. must now match the active HTTP client (`httpx2` by default).

! `CodeExecutionResultStep.result` changed from `string` to the `CodeExecutionResultStepResult` union; code that reads `result` as a plain string must now check which variant is present and extract the string from it.

WAS THIS USEFUL? 👍 👎

Tracer Cloud OpenSRE

Feedback ⓘ v0.1.2026.11.6 NOTES ↗

OpenSRE is an open-source toolkit for building AI SRE agents.

OpenSRE's GitHub CLI reader now batches reads into one call and skips discovery, cutting API overhead.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.1.2026.11.6 https://github.com/Tracer-Cloud/opensre.git
# already have the repo? check out this version:
$ git checkout v0.1.2026.11.6
```

NEW INTEGRATION Neon read-only integration and remote-sync provider

---

WAS THIS USEFUL? 👍 👎

feder-cr gitai

Feedback ⓘ v0.4.2 NOTES ↗ RANK

Git designed for AI agents.

> Across v0.4.1 and v0.4.2, gitai added a repair command for conflicting storage refs, Postgres connection-pool tuning, write-lock timing metrics, a dry-run policy validator, and signal-based hot reload of running policies.

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

01 Repair command for conflicting directory/file refs NEW

90

The new `graft storage delete-conflicting-ref REF EXPECTED_TARGET` command (backed by RPC `DeleteConflictingReference`, authoritative mode only) deletes one side of a conflicting directory/file reference pair only if it still holds the expected value, and applies the deletion as an ordinary changeset propagated across all nodes.

Remove one side of a refs/heads/a vs refs/heads/a/b conflict that is blocking node catch-up in authoritative mode.

```
$ graft storage delete-conflicting-ref refs/heads/a/b <expected-sha>
```

— Exact command, RPC name, and safety semantics given v0.4.2

02 Policy dry-run validator command NEW

90

`graft policy check FILE` with the `-repo` flag validates a policy file and replays recent changesets against it, exiting 0 when nothing is denied, 1 when updates would be denied, and 2 when the file is invalid.

Test a candidate policy file against a live repository's recent changesets before deploying it, so a bad policy never takes effect.

```
$ graft policy check ./new-policy.toml -repo /srv/gitai/repos/myrepo
```

— Exact command, flag, and exit codes with usage example v0.4.1

03 Postgres connection pool tuning in storage config NEW

88

Four new config keys under `[graft.storage]` tune Postgres pooling: `database_max_open_conns` (default 32), `database_max_idle_conns` (defaults to max open), `database_conn_max_idle_time` (default 5m), and `database_conn_max_lifetime` (default 30m). Setting any of these while using SQLite is a startup error.

Cap Postgres connections across a three-node cluster so the total stays below the server's max_connections limit, with longer idle-connection reuse to reduce churn.

```toml
database_max_open_conns = 20
database_max_idle_conns = 20
database_conn_max_idle_time = '10m'
database_conn_max_lifetime = '30m'
```

— All four keys named with defaults and a config example v0.4.2

04 **Hot policy reload via SIGUSR1** NEW 71

Sending `SIGUSR1` reloads the standalone policy from its configuration file at runtime; the swap is atomic, and an invalid file leaves the old policy running.

Apply an updated policy to a running gitai server without downtime or dropped connections.

```
$ kill -SIGUSR1 $(pidof gitaly)
```

— Named signal and fallback behavior with example command v0.4.1

THINNER COVERAGE BELOW

05 **Write-lock timing metric for storage steps** NEW 59

gitai now exposes `graft_storage_prepared_step_seconds{kind,step}` to split time spent under the repository write lock into record, local read, pack, commit, and durability steps.

— Named metric and labels but no usage example given v0.4.2

06 **Bundled git version bump to v2.56.0** IMPROVED 30

The default bundled git release is now v2.56.0, up from 2.55.0-rc1.

— Exact before/after version but no further detail v0.4.2

07 **Git conformance harness in CI** NEW 22

gitai now runs a conformance harness comparing stock git against gitai in every storage mode as a CI step.

— Brief mention with no detail on scope or output v0.4.2

WAS THIS USEFUL? 👍 👎

Anthropic TypeScript SDK Feedback ⓘ sdk-v1.6.0

NOTES ↗ RANK

The Anthropic TypeScript SDK provides typed JavaScript and TypeScript clients for Claude APIs, streaming, and tool use.

> Anthropic's TypeScript SDK adds new error codes for rate and key-validation failures and introduces organization-wide member sign-out via the API.

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

01 **New error codes for ramp limits and External Key validation**

NEW 35

The SDK adds an error code for ramp limit errors so callers can distinguish rate/ramp throttling from other API errors, and adds error codes for External Key validation failures so callers can detect and handle third-party key problems explicitly.

— Describes purpose but no exact code names or values given sdk-v1.6.0

02 **Organization-wide member sign-out via API** NEW 26

Adds support for signing organization members out everywhere via the API.

— No endpoint, method or method name specified sdk-v1.6.0

WAS THIS USEFUL? 👍 👎

AWS Amazon Bedrock AgentCore

Feedback ⓘ

Gateway: Get cited web answers with the Web Answer connector (2026-11-05)

BLOG ↗ RANK

Amazon Bedrock AgentCore provides managed agent runtimes, memory, identity, gateways, observability, and evaluation services.

> Amazon Bedrock AgentCore Gateway adds a Web Answer connector that returns synthesized, cited answers from live web sources via a single MCP tool call.

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

01 **Web Answer connector in Gateway** NEW 55

Amazon Bedrock AgentCore Gateway now offers a Web Answer connector, available in US East (N. Virginia), that lets an agent send a question and receive a synthesized answer with citations to current web sources from a single MCP tool call.

— Names region and mechanism but no API/flag details

Gateway: Get cited web answers with the…

WAS THIS USEFUL? 👍 👎

Anthropic Python SDK

Feedback ⓘ v2.6.0 NOTES ↗ RANK

The Anthropic Python SDK provides typed Python clients for Claude APIs, streaming responses, and tool use.

> Anthropic Python SDK v2.6.0 adds distinct API error codes for ramp-limit throttling and External Key validation failures, plus a new organization-wide sign-out operation.

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

01 **New error codes for ramp limits and External Key validation**

NEW 37

The API now exposes a distinct error code for ramp limit errors, letting applications detect and handle rate-ramp throttling separately from other error types, and a separate error code for External Key validation failures so applications can identify and handle key-validation errors programmatically.

— Names two new error codes but no exact identifiers or handling code v2.6.0

02 **Organization-wide sign-out operation** NEW 33

The API now supports signing all members of an organization out everywhere in a single operation.

— States the capability but no endpoint name or usage shown v2.6.0

WAS THIS USEFUL? 👍 👎

OpenAI TypeScript SDK

Feedback ⓘ v7.52.0 NOTES ↗

RANK

The OpenAI TypeScript SDK provides typed JavaScript and TypeScript clients for OpenAI APIs and streaming responses.

> OpenAI TypeScript SDK now lets callers set the model per-request on the Responses API.

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

01 **Per-request model option on Responses API** NEW 60

The Responses API now accepts a `model` option passed directly to client.responses.create(), allowing the model to be specified per request.

Specify which model to use when creating a response, now that the Responses API accepts a `model` field directly.

typescript

```typescript
const response = await client.responses.create({
  model: 'gpt-4o',
  input: 'Summarize this document.',
});
console.log(response.output_text);
```

— Names exact method and option, but no further mechanism detail. v7.52.0

WAS THIS USEFUL? 👍 👎

OpenAI Python SDK

Feedback ⓘ v3.47.0 NOTES ↗ RANK

The OpenAI Python SDK provides typed Python clients for OpenAI model APIs, streaming, and asynchronous requests.

> The OpenAI Python SDK now lets callers specify a model directly on the Responses API.

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

01 **Model option on Responses API** NEW 60

The `model` option can now be passed to client.responses.create() to specify which model handles the request, e.g. `model="gpt-5.5"`.

Select a specific model for a Responses API call now that the model option is supported.

python

```python
from openai import OpenAI

client = OpenAI()

response = client.responses.create(
    model="gpt-5.5",
    input="Explain workload identity in one paragraph.",
)
print(response.output_text)
```

— Names exact method and parameter with runnable example. v3.47.0

WAS THIS USEFUL? 👍 👎

Parakeet ASR Server

Feedback ⓘ v0.9.0 NOTES ↗ RANK

Parakeet ASR Server transcribes audio on CPUs or GPUs using NVIDIA Parakeet ONNX models through a Whisper-compatible API.

> Parakeet ASR Server v0.9.0 adds opt-in speaker diarization, real sentence-level timestamps for subtitle formats, and two mel-frontend accuracy fixes — one of which changes the default chunk window and will reprocess long audio differently.

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

01 Opt-in speaker diarization NEW

88

Enable diarization by passing `-enable-diarization` at startup or setting `PARAKEET_ENABLE_DIARIZATION=true`, then add `diarize=true` to individual transcription requests. Parakeet runs NVIDIA Nemotron-3-Diarization as a second ONNX model alongside the main transcription model and returns speaker-tagged word turns.

Transcribe a multi-speaker meeting recording and get speaker-tagged turns attributed per word.

```
$ ./parakeet -enable-diarization
# then in a separate request:
curl -F "file=@meeting.mp3" -F "diarize=true"
http://localhost:5092/v1/audio/transcriptions
```

— Names flag, env var, request param and model used v0.9.0

02 Real sentence-level timestamps in subtitle formats IMPROVED

77

The `verbose_json`, `srt`, and `vtt` response formats now return real sentence-level segments with accurate start/end timestamps at 80 ms granularity, honest `avg_logprob`, and `compression_ratio`, producing one cue per segment for subtitle outputs.

Generate a subtitle file with real sentence-level cues and accurate timestamps from a long recording.

```
$ curl -F "file=@meeting.mp3" -F "response_format=srt"
http://localhost:5092/v1/audio/transcriptions
```

— Names formats and fields but limited mechanism detail v0.9.0

03 **Per-window mel normalization and smaller default chunk window** 75

BREAKING

Mel features are now normalized per window instead of across the entire file, reducing word error rate on long recordings from 5.4% to 3.4% on a 7-minute English test. The default chunk window shrank from 300s to 60s with 15s overlap; pipelines relying on the previous default will now process long audio in smaller chunks.

— Gives before/after numbers and a breaking default change v0.9.0

THINNER COVERAGE BELOW

04 **NeMo-faithful mel filter bank** IMPROVED 43

The mel filter bank now matches NVIDIA NeMo preprocessing, using the same Slaney filter bank, pre-emphasis, and normalization, closing accuracy drift for non-English languages.

— Explains mechanism but no reproducible steps v0.9.0

└──▷ BREAKING ON UPGRADE

! The default chunk window changed from 300s to 60s; pipelines that relied on the previous default window size will process long audio in smaller chunks.

WAS THIS USEFUL? 👍 👎

MotleyAI SLayer

Feedback ⓘ v1.7.0 NOTES ↗ RANK

SLayer is an embeddable semantic layer that lets AI agents and people query warehouse data through shared metric definitions.

> SLayer v1.7.0 adds query support for array/list columns with new functions and tooling to re-ingest and validate type changes, and switches Postgres connections to psycopg v3 by default.

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

01 **Query functions for list/array columns** NEW 87

Array columns can now be queried as typed lists using contains(skills, 'sql'), contains_any(...), contains_all(skills, ['go', 'sql']), and length(skills) across Postgres, DuckDB, BigQuery, Snowflake, ClickHouse, Trino, and Presto. Individual list elements can be used as a dimension or filter via `skills.value` without double-counting, and list values can be aggregated per group with the new collect(x) aggregation.

Count users per skill from an array column without double-counting other rows in the query.

```json
{"source_model": "users", "measures": ["count(*)"], "dimensions": ["skills.value"]}
```

Filter to users who know both Go and SQL using the new list predicate.

```json
{"source_model": "users", "filters": ["contains_all(skills, ['go', 'sql'])"], "measures": ["count(*)"]}
```

— Names every function and engine with runnable query examples v1.7.0

02 **Postgres defaults to psycopg v3 driver** BREAKING 80

Postgres datasources now connect with psycopg (v3) by default when no driver is named in the URL; install it via the `postgres` extra or name an explicit driver in the URL (e.g. `postgresql+psycopg2://...`). Connecting with psycopg2, asyncpg, or aiomysql now warns that their extras will stop installing them in 2.0.

— Exact URL syntax and extra name given, but no runnable command v1.7.0

03 **Re-ingest and validate type changes on array columns** NEW 78

Array columns previously saved as text can be re-ingested to convert them to lists, with the re-ingest reporting every saved measure, derived column, model filter, and saved query the change breaks. Running `slayer validate-models --force-clean` leaves a column whose type changed, and everything built on it, in place instead of deleting it.

Leave changed columns and their dependents in place when validating models, rather than deleting them.

```
$ slayer validate-models --force-clean
```

— Names the exact flag and describes the breakage-reporting mechanism v1.7.0

---

04 **SQL client method renamed to get_result_column_types**

DEPRECATED 60

Custom SQL clients implementing `get_column_types` must rename it to `get_result_column_types`; a deprecation warning is issued and the old name still works until 2.0.

— Names old and new method but limited mechanism detail v1.7.0

└──▷ **BREAKING ON UPGRADE**

! A Postgres URL or config that names no driver now connects with psycopg (v3) instead of psycopg2 and asyncpg; if your environment has psycopg2 or asyncpg installed but not psycopg, connections will fail unless you name the driver explicitly in the URL (e.g. `postgresql+psycopg2://...`) or install psycopg.

! Connecting with psycopg2, asyncpg, or aiomysql now warns that their extras will stop installing them in 2.0.

! Custom SQL clients implementing `get_column_types` must rename it to `get_result_column_types` (a deprecation warning is issued; the old name still works until 2.0).

WAS THIS USEFUL? 👍 👎

◆ Inference Providers & GPU Clouds

Eigen Labs Darkbloom

Feedback ⓘ v0.10.10 NOTES ↗

CODE ↗ RANK

Darkbloom is a for the innovators pushing the frontier.

> Darkbloom introduced a waitlist-gated approval system for cluster and Sandbox access, added GPU memory tuning and external-disk support for Apple Silicon pairs, and sped up updates with automatic version parity between paired Macs.

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

01 GPU memory limit for Apple Silicon pairs NEW

95

New `darkbloom cluster gpu-limit` lets a 128 GB or larger Apple Silicon Mac reserve its GPU memory less 16 GiB (112 GiB on a 128 GB Mac), leaving room for longer requests on a pair; `darkbloom cluster gpu-limit --undo` restores macOS's default.

Free up GPU memory on a 128 GB Mac so a pair can accept longer prompts without running out of memory.

```
$ darkbloom cluster gpu-limit
```

— Exact command, flag, and memory numbers with runnable example v0.10.9

02 Waitlist system for cluster and Sandbox access BREAKING

90

New `darkbloom waitlist` checks approval status across all features, or joins a specific list with `darkbloom waitlist cluster` or `darkbloom waitlist sandbox`; add `--json` for the coordinator's raw answer. A first-time `darkbloom sandbox setup` and a first `darkbloom cluster` on an unpaired Mac now require an approved waitlist entry before proceeding and exit 1 if none exists, where previously they proceeded unconditionally; pass `--yes` on the sandbox path to join the waitlist without an interactive prompt.

See where your account stands on every waitlist before attempting a new cluster or Sandbox setup.

```
$ darkbloom waitlist --json
```

Join the cluster waitlist so a new two-Mac pair can eventually be approved to serve through Darkbloom.

```
$ darkbloom waitlist cluster
```

Set up a new Sandbox and automatically join the sandbox waitlist in one step without an interactive prompt.

```
$ darkbloom sandbox setup --yes
```

— Names every command and flag with exact gating behaviour and runnable examples v0.10.10

03 **External disk support for Sandbox VMs** NEW 80

New `--disk <volume>` flag (repeatable) on `darkbloom sandbox setup` lets owners place VM images and job disks on external APFS drives instead of internal storage.

Direct sandbox VM images and job disks onto an external APFS drive to save internal SSD space and wear.

```
$ darkbloom sandbox setup --disk /Volumes/ExternalSSD
```

— Named flag with concrete path example, mechanism stated v0.10.9

04 **Faster updates and automatic version parity** IMPROVED 75

`darkbloom update` now ends as soon as the local Mac runs the new release instead of waiting a fixed timeout, and the leading Mac in a pair checks every 5 seconds instead of every minute. Paired Macs now keep version parity automatically: when one Mac runs a newer release, the other's cluster service detects the difference within a minute and updates immediately instead of waiting for the next update slot.

— Names the command and timing changes but update is automatic, no user action v0.10.9

05 **Solo serving flag for unpaired Macs** NEW 65

New `darkbloom start --solo` flag stops any leftover pair part and serves on a single Mac alone, replacing the need to manually clear pair state.

Restart solo serving on a Mac whose pair is no longer running, without manually clearing pair state.



— Named flag with clear runnable example v0.10.9

06 **Not-serving status reporting for pairs** IMPROVED 65

`darkbloom cluster status` now reports `not_serving` with exit code 3 when a pair set to serve through Darkbloom is not accepted, and both screens show ✗ Not serving: not offered to Darkbloom (...).

— Exact exit code and command named but no runnable example given v0.10.9

THINNER COVERAGE BELOW

07 **Local clustering without a waitlist or coordinator** NEW 55

Run `darkbloom cluster local` to serve across your own Macs with no waitlist check and no network coordinator required.

— Command named but mechanism beyond 'no coordinator' is thin v0.10.10

08 **Model unavailability explanation on pairs** IMPROVED 20

A model approved for clusters that cannot be offered on the current pair now shows a one-line explanation above the model list instead of silently disappearing.

— Only a brief prose description, no named surface or mechanism v0.10.9

└──▷ **BREAKING ON UPGRADE**

! A new (first-time) `darkbloom sandbox setup` and a first `darkbloom cluster` on an unpaired Mac now exit 1 unless the account holds an approved waitlist entry — where previously they proceeded unconditionally.

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

MLX Serve

Feedback ⓘ v26.11.2 NOTES ↗ RANK

The mlx-serve runtime provides Apple Silicon inference with OpenAI- and Anthropic-compatible APIs, chat, agent mode, and tool calling.

> mlx-serve's biggest release of the window adds dual-Mac tensor parallelism for joint inference across two Thunderbolt 5 Macs and extends GGUF support to the faster MLX engine, alongside conversation checkpoint caching, per-request activation steering, and an MLX engine bump that requires rebuilding existing plugins.

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

01 GGUF support on the MLX engine IMPROVED

95

Run DeepSeek-V4.1 GGUF file sets on the MLX engine using `--mlx-gguf`, gaining DSpark drafting, image input, and prompt throughput around 3,300 tokens per second on M5 Ultra — about 2.5x the previous rate. More GGUF families (Qwen3, Flint, Muse Glimmer) now also take the MLX engine path with `--mlx-gguf`, with faster IQ kernels.

Run a DeepSeek-V4.1 GGUF pack on the faster MLX engine to get DSpark drafting, image input, and ~2.5x prompt throughput.

```
$ mlx-serve --mlx-gguf
```

— Runnable flag with throughput numbers and named families v26.11.2

02 Dual-Mac tensor parallelism NEW

85

Split a model across two Thunderbolt 5 Macs for joint token generation using `--dist tensor`, delivering up to 57% faster output on two M3 Ultras.

Distribute a large model across two Thunderbolt 5 Macs to get joint token generation and up to 57% faster output.

```
$ mlx-serve --dist tensor
```

— Runnable flag with concrete hardware benchmark v26.11.2

03 MLX 0.32.5 upgrade requires plugin rebuild BREAKING

65

mlx-serve now runs on MLX 0.32.5; outside plugins built for 26.11.1 show `Needs rebuild` in Models › Plugins and can be recompiled with `Rebuild all`.

— Exact UI path and remediation action given v26.11.2

THINNER COVERAGE BELOW

04 **MCP server settings preserved in mcp.json** IMPROVED

55

MCP server settings edited by hand in `mcp.json` (command, environment, folder) are now preserved when toggling a server on or off in the marketplace.

— Names config file and fields directly v26.11.2

05 **Conversation checkpoint caching** NEW

50

Long conversations on stateful models (DeepSeek-V4.1 and others) now keep earlier checkpoints so editing or forking an older message resumes in a fraction of a second rather than re-reading the full chat.

— Describes mechanism but no exact command or config v26.11.2

06 **Higher context length for GGUFs on DeepSeek engine** IMPROVED

45

GGUFs on the built-in DeepSeek engine now open at the context length they were trained for instead of being capped at 32k.

— Clear before/after but no actionable flag given v26.11.2

07 **Models stay resident in memory** IMPROVED

40

Models now stay resident in memory between requests instead of being evicted by macOS pressure, preventing freezes on Macs running models close to their memory limit.

— No config key or flag to control this behavior v26.11.2

08 **Plugin access to Neural Engine and SSD streaming** NEW

30

Outside plugins can now use the Neural Engine and stream weights from SSD with their own memory budget.

— Thin description with no API or flag named v26.11.2

09 **Per-request activation steering** NEW

30

A server can now steer a model per request by removing a direction from its activations.

— Mechanism named but no API or flag given v26.11.2

└──▷ ALSO FROM THESE RELEASES

└──▷ **BREAKING ON UPGRADE**

! Outside plugins built for v26.11.1 no longer load automatically after upgrading to MLX 0.32.5; they show `Needs rebuild` in Models › Plugins and must be recompiled.

WAS THIS USEFUL?  

Ollama

Feedback ⓘ v0.40.15 NOTES ↗ CODE ↗ RANK

Ollama downloads, serves, and chats with local language models through a CLI and REST API.

> Ollama's v0.40.15 release adds two new local models (AutoTrust decision classifiers and MiMo V2.6), removes the single-request limit for recurrent and vision models, and rounds out several `ollama launch` UX improvements.

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

01 AutoTrust decision models (autotrust-jev) NEW

75

Run the `autotrust-jev` family (9b, 35b, 27b-vl) with `ollama run autotrust-jev`; the 35b and 27b-vl sizes accept images alongside text, and the model returns yes/no probabilities and ratings for questions supplied via the `--ask` flag.

Query an AutoTrust decision model to classify a customer support message — useful for routing or triage pipelines that need a structured yes/no verdict.

```
$ ollama run autotrust-jev "A customer says their card was charged twice." --ask "Is this a billing issue?"
```

— Names model variants and --ask flag with a runnable example v0.40.15

THINNER COVERAGE BELOW

02 Parallel requests for recurrent and vision models IMPROVED

55

Recurrent and vision GGUF models now serve more than one request at a time, removing the previous `num_parallel=1` limit that applied to those architectures.

— Explains before/after but no command to invoke v0.40.15

03 Ollama launch UX improvements IMPROVED

55

`ollama launch` now supports Kimi Code CLI, the successor to kimi-cli, and offers to migrate existing kimi-cli setups; it also fills in Pi's model list automatically so models can be picked inside Pi, and Claude Desktop's model picker now shows Ollama's actual model names instead of route IDs when launched via `ollama launch`. Separately, running `ollama` with no arguments now opens a welcome launcher instead of printing help text.

— Names all four launcher surfaces but no deep mechanism v0.40.15

04 MiMo V2.6 code and agent model NEW

50

Run the `mimo-v2.6` 9B code and agent model locally with `ollama run mimo-v2.6`.

Run the MiMo V2.6 code and agent model locally for coding tasks or agentic workflows.

```
$ ollama run mimo-v2.6
```

— Runnable command but no detail on capabilities v0.40.15

05 **Long-context YaRN scaling for Qwen3** IMPROVED 30

Qwen3 and Qwen3.5 models now support long-context YaRN scaling.

— Names the models and technique but no configuration details v0.40.15

WAS THIS USEFUL? 👍 👎

◆ Inference Serving & Optimization

LMCache

Feedback ⓘ v0.5.7 NOTES ↗ RANK

LMCache reuses and shares language-model key-value caches across requests and serving instances to reduce repeated prefill work.

> LMCache v0.5.7 adds opt-in gRPC transports for multi-process serving, broadens accelerator support to Biren, AWS Trainium/Neuron, Cambricon and Ascend, and introduces engine-driven multi-group KV transfers alongside new prefetch-affinity and shared-memory controls; official Docker images now run as a non-root user, a breaking change for upgraders.

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

01 **MP server transports and module extensibility** NEW — 73

Expands multi-process serving with opt-in `grpc.aio` and `pygrpc-native` transports, automatic server startup for `grpc://` endpoints, and `grpc+unix` socket support, and adds MP server-module client-affinity calls and pluggable transfer hooks, allows MP server modules to replace built-in modules, and emits per-request prompt trace spans from the MP server.

— Names transports and endpoint protocols but no exact config key Describes capabilities without naming specific APIs

v0.5.7

02 **Prefetch affinity and admission controls** NEW — 70

Adds request tags for L1 managers and L2 adapters to enable affinity and prefetch filtering, adds L1/L2 allowlists when submitting prefetch requests to the storage manager, and adds dynamic prefetch admission to the prefetch controller.

— Names components involved but no config keys given v0.5.7

03 **Engine-driven multi-group KV transfers** NEW — 65

Adds engine-driven multi-group transfers through `KVLayerGroupsManager`, and extends these transfers to support compressed KV groups for DeepSeek V4.

— Names the manager class and target model, no usage shown v0.5.7

04 **Byte-sized fs_native read depth** IMPROVED — 60

Decouples `fs_native` read depth from worker count and sizes it in bytes instead.

— Names the surface changed but no exact parameter name v0.5.7

05 **Platform and accelerator support expansion** NEW — 60

Adds Python 3.14 support, extends accelerator coverage with Biren SUPA, AWS Trainium/Neuron for MP KV caching, Cambricon MLU, and Mamba models on Ascend, and supports the blocks-first KV layout in `vLLM-RBLN`.

— Lists many named platforms but no setup steps v0.5.7

06 **Shared and device-addressable KV memory** NEW 60

Advances shared and device-addressable KV memory with a shared-L1 memory orchestrator, Device-DAX/CXL L1 offsets, and range-based DAX L2 ownership.

— Names mechanisms but omits configuration or usage detail v0.5.7

THINNER COVERAGE BELOW

07 **Non-root Docker images** BREAKING 55

Official Docker images now run as a non-root user — on upgrade, verify that writable paths and mounted-volume permissions are accessible to the new user.

— Clear upgrade action but no specific UID or paths named v0.5.7

08 **KV events in MultiConnector P/D flows** NEW 45

Supports KV events in `MultiConnector` prefill/decode (P/D) flows.

— Names the connector but no event detail or usage v0.5.7

09 **Coordinator /status endpoint and Grafana dashboard** NEW 35

Adds a coordinator `/status` endpoint and a Grafana dashboard for monitoring.

— Mentioned only briefly, no fields or dashboard contents given v0.5.7

10 **Per-request KV-load versus recompute policy** NEW 33

Adds a configurable per-request policy for choosing between loading cached KV data or recomputing it.

— No flag, key, or mechanism named v0.5.7

└──▷ **BREAKING ON UPGRADE**

! Official Docker images now run as a non-root user — verify that writable paths and mounted-volume permissions are accessible to the new user on upgrade.

WAS THIS USEFUL? 👍 👎

NVIDIA Dynamo

Feedback ⓘ v1.6.2 NOTES ↗ RANK

Dynamo is a datacenter-scale distributed inference serving framework with disaggregated prefill and decode, KV-cache-aware routing, and multi-engine support.

> NVIDIA Dynamo v1.6.2 adds an opt-in garbage-collection freeze option for SGLang workers to cut streaming stalls, alongside three breaking changes to the SGLang/Planner base images and the Mooncake transfer engine version floor.

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

01 **GC freeze flag for SGLang workers** NEW 75

Reduce streaming stalls in long-running SGLang decode and aggregated workers by enabling garbage-collection freeze after startup, via `--freeze-gc-after-init` or the environment variable `DYN_SGL_FREEZE_GC_AFTER_INIT`; off by default.

– Names exact flag and env var, behavior and default clearly stated. v1.6.2

02 **Planner image build requires newer runtime pip** BREAKING 60

The Planner image build now fails if the runtime base bundles an older pip, following the move to `python:3.13.18-slim` for the build and `distroless Python 3.13-v4.1.10` for the runtime.

– Names exact base images but no fix instructions given. v1.6.2

THINNER COVERAGE BELOW

03 **cuda-gdb loses Python scripting in SGLang image** BREAKING 53

`cuda-gdb` in the SGLang image no longer supports Python scripting, because its Python front end links the Ubuntu `libpython3.12` that the move to CPython 3.12.16 removes.

– Explains cause but gives no remediation step. v1.6.2

04 **Mooncake transfer engine version floor raised** BREAKING 40

The Mooncake transfer engine floor in the SGLang and vLLM images is raised from `0.3.13.post1` to `0.3.14`.

– Version numbers given but no explanation of impact or mechanism. v1.6.2

└──▷ BREAKING ON UPGRADE

! `cuda-gdb` in the SGLang image no longer supports Python scripting, because its Python front end links the Ubuntu `libpython3.12` that the move to CPython 3.12.16 removes.

! The Planner image build now fails if the runtime base bundles an older pip, following the move to `python:3.13.18-slim` for the build and `distroless Python 3.13-v4.1.10` for the runtime.

! The Mooncake transfer engine floor in the SGLang and vLLM images is raised from `0.3.13.post1` to `0.3.14`.

WAS THIS USEFUL?  

Niko1221 Strata

Feedback ⓘ v0.2.11 NOTES ↗

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 adds chat-to-file saving, faster NVIDIA K-quant prompt kernels (up to +38%) and larger prompt chunks (32768 tokens), and auto split-engine selection for multi-Radeon setups.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.2.11 https://github.com/Niko1221/Strata.git
# already have the repo? check out this version:
$ git checkout v0.2.11
```

Chats are now saved as files.

NVIDIA prompt speed for K-quant files (Q4/Q5/ `Q6_K_XL`) now uses llama.cpp's MMQ kernels by default on cards of compute capability 8.0 and up, raising prompt throughput by up to 37% on an RTX PRO 6000.

Bare `--prefill auto` now reads prompts in chunks of up to 32768 tokens (up from 8192) on NVIDIA cards of compute capability 8.0 and up, increasing prompt speed on long contexts by up to 43% on an RTX A6000.

Four Radeon AI PRO R9700 cards with the all-cards split engine now auto-selects the split engine for `IQ3_S` sets via a plain `strata launch --gpu 0,1,2,3` , raising single-user decode from 104 to 187 tokens/s and 8-user combined from 321 to 365 tokens/s.

Four Radeon AI PRO R9700 cards now use 8 request slots by default (up from 6), and prompt sums route through other cards' memory so `libreccl` is no longer required.

NEW SUPPORTED PLATFORM Unsloth Qwen3.8-Flash-Next GGUF variants (Dynamic v3.1)

NEW SUPPORTED PLATFORM Q5_0 expert format support

NEW SUPPORTED PLATFORM Mac support for Unsloth Qwen3.8-Flash-Next sets

WAS THIS USEFUL? 👍 👎

◆ AI / LLM Gateways

Mozilla AI Otari

Feedback ⓘ v0.39.0 NOTES ↗ RANK

Open-source, OpenAI-compatible LLM gateway you run yourself. One endpoint for 40+ providers, with virtual keys, budgets, and usage tracking.

> Mozilla AI Otari's v0.39.0 release adds a direct gateway fetch endpoint, per-workspace provider restrictions, zero data retention tracking, and new metering and cost-reporting for video and image generation.

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

01 **Direct fetch endpoint bypassing model calls** NEW 75

Call the new `POST /api/v1/fetch` endpoint to execute a direct fetch from the gateway without going through a model.

– Named endpoint with method and path, but mechanism lightly described v0.39.0

02 **Video generation metering and job tracking** NEW 65

Otari now meters and prices video generation jobs by the second and adds a videos endpoint to record and retrieve video jobs that callers start.

– Mechanism and new endpoint described but exact path not given v0.39.0

03 **Configurable rates in pricing API** IMPROVED 60

The pricing API and config now accept and return rates, and the catalog's cost block seeds into a hosted model's rate.

– Mechanism named but no exact config key or endpoint path v0.39.0

THINNER COVERAGE BELOW

04 **Per-workspace provider and model restrictions** NEW 50

Organization admins can restrict which providers and models each workspace may use.

– Describes control but no named config key or API v0.39.0

05 **Zero data retention declarations per provider** NEW 50

Operators can declare zero data retention for a provider so the guarantee is tracked and surfaced at the operator level.

– Explains behavior but no named surface or exact mechanism v0.39.0

06 **Inline settled cost in image responses** IMPROVED 35

Image generation and edit responses now return the settled cost inline.

— Brief description with no named field or endpoint `v0.39.0`

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

pymilvus

Feedback ⓘ v2.6.18 NOTES ↗ RANK

PyMilvus is the Python SDK for Milvus, an open-source vector database for AI applications.

> pymilvus v2.6.18 adds row-level security enforcement, configurable Volume upload policies, idempotent bulk import support via custom headers, and new async RPC methods on AsyncMilvusClient.

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

01 Configurable Volume upload policy NEW

90

Control Volume upload behavior by setting `upload_policy=UploadPolicy.OVERWRITE` on VolumeFileManager.upload_file_to_volume() to replace matching objects, or rely on the new default which skips existing objects whose key and byte size match. Paginated listings and bounded concurrency are also now configurable for large batches.

Force-overwrite an existing Volume object instead of skipping it when the key and size already match.

python

```python
from pymilvus.bulk_writer import VolumeFileManager, UploadPolicy

manager = VolumeFileManager(url="http://localhost:19530",
volume_name="my_volume")
manager.upload_file_to_volume(
    local_path="/data/vectors.parquet",
    upload_policy=UploadPolicy.OVERWRITE
)
```

— Names exact method, flag and default, with runnable example v2.6.18

02 Custom headers and idempotency key for bulk import NEW

90

Pass custom HTTP headers to bulk_import(), get_import_progress(), and list_import_jobs(); the SDK now exports `IDEMPOTENCY_KEY_HEADER` for use with servers that support idempotent bulk imports, so retrying the same request does not create duplicate import jobs.

Submit a bulk import with an idempotency key so that retrying the same request does not create duplicate import jobs.

```python
from pymilvus import IDEMPOTENCY_KEY_HEADER

bulk_import(
    url="http://localhost:19530",
    collection_name="my_collection",
    files=[["s3://bucket/data.parquet"]],
    headers={IDEMPOTENCY_KEY_HEADER: "import-job-2024-06-01-001"}
)
```

— Names exact functions and header constant with runnable example v2.6.18

03 **Row-level security policy enforcement** NEW 70

Enforce row-level security at the server by passing `rls_principal` to read and write operations on `MilvusClient` and `AsyncMilvusClient`, and manage RLS policies and principal tags through new policy management APIs. This requires a Milvus deployment with RLS support.

— Names exact param and client classes but no runnable example v2.6.18

04 **New async RPC methods on AsyncMilvusClient** NEW 60

Call AsyncMilvusClient.flush_all(), get_compaction_plans(), and update_replicate_configuration() asynchronously via new async RPC handlers.

— Names exact methods but no usage example or behavior detail v2.6.18

WAS THIS USEFUL? 👍 👎

knowhere

Feedback ⓘ v1.2.45 NOTES ↗ RANK

Knowhere extracts, parses, and outputs structured chunks ready for AI Agents and RAG.

> Knowhere improved V2 full ingest handling of short scanned PDFs lacking a table of contents, now producing page citation images instead of empty chunks.

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

01 **Page citation images for short scanned PDFs** IMPROVED

53

During V2 full ingest, short scanned PDFs without a table of contents now produce page citation images instead of an empty whole-document chunk; short text PDFs without a table of contents continue to return their original page text unchanged.

— Describes behavior change and conditions but no config or API surface. v1.2.45

WAS THIS USEFUL? 👍 👎

aegis-stack

Feedback ⓘ v0.19.0 NOTES ↗ RANK

A production-ready FastAPI platform with modular components and a built-in control plane.

> aegis-stack's v0.19.0 release adds a JSON-based `aegis menu` discovery command, new chat-mount controls for public access, call bounding, and agent-navigable pages, alongside several breaking changes that tighten deploy, init, and health-check behavior.

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

01 `aegis menu` JSON component discovery NEW

90

The new `aegis menu` command prints every component and service, their bracket options, and dependency rules as JSON; pass a selection to get back the `init` command that builds it or the rules that block it. Each GitHub release now also ships a `menu.json`.

Discover what brackets and dependencies a component accepts before wiring up a new project.

```
$ uvx aegis-stack menu
```

— Names command, output format, and ships a runnable example v0.19.0

02 New chat mount controls NEW

75

Chat mounts gain new controls: `ChatSurface.calls` toggles whether unsigned visitors may call a mount while hiding engine and cost details from public callers; `longest_call`, `idle_seconds`, `call_check`, and `on_call_end` bound call duration and idle time so calls end gracefully and the conversation continues in text; and `ChatSurface.pages` lets you declare agent-navigable pages, each listing the parts an agent may scroll to, highlight, or act on, with only a declared page exposing the `page_action` tool.

— Names the config key and its effect, no usage example Four config fields named, mechanism explained, no example shown Config surface and gated tool named, mechanism described

v0.19.0

03 Health API and CLI changes (BREAKING) BREAKING

75

`health probe` now exits 1 when the API does not answer or a component is unhealthy, where it previously passed whenever the process was up; the health API's `status` field on `/health/detailed` and `/health/dashboard` now also returns `warning` or `info` and those endpoints now include host metrics; and `health`

`status` now reads `/health/detailed`, which from a worker or scheduler container in production returns a 401 or 404, so it should be run from the webserver container instead.

— Exact command and before/after behavior stated Two endpoints and new field values named explicitly Exact endpoint, failure codes, and correct container named

v0.19.0

---

04 **Stricter `aegis init` bracket and scheduler validation** 70

BREAKING

`aegis init` now refuses bracket syntax it previously accepted (`database[mysql]`, `redis[x]`) and rejects a persistent scheduler paired with a mismatched database engine.

— Specific rejected syntax named but no migration command given v0.19.0

---

05 **Deploy stops on backup failure** BREAKING 65

A deploy whose backup fails now stops instead of continuing; pass `--no-backup` to proceed without a backup.

— Clear before/after and an exact flag to opt out v0.19.0

---

06 **Hide raw AI call details in chat step trail** NEW 60

Turning off `AI_CHAT_SHOW_CALLS` hides raw AI call details from the chat step trail, showing a plain description of what each step is doing instead.

— Env var named, behavior change described, no example v0.19.0

└──▷ **BREAKING ON UPGRADE**

! A deploy whose backup fails now stops instead of continuing — use `--no-backup` to proceed without a backup.

! `aegis init` now refuses bracket syntax it previously accepted (`database[mysql]`, `redis[x]`) and rejects a persistent scheduler paired with a mismatched database engine.

! `health probe` now exits 1 when the API does not answer or a component is unhealthy; it previously passed whenever the process was up. Anything using it as a gate may start failing.

! The health API's `status` field on `/health/detailed` and `/health/dashboard` now returns `warning` or `info` in addition to its previous values, and those endpoints now include host metrics.

! `health status` now reads `/health/detailed`; from a worker or scheduler container in production this returns a 401 or 404 — run it from the webserver container instead.

WAS THIS USEFUL?  

GrowGraph OntoCast

Feedback ⓘ v0.9.0 CODE ↗ RANK

OntoCast generates RDF knowledge graphs from documents by co-evolving domain ontologies and fact graphs with an agentic LLM pipeline.

> OntoCast's v0.9.0 release adds an LLM-based coreference resolution step to fact merging, embedding-based heading labelling, table-of-contents detection, and detailed per-unit routing diagnostics in the chunking pipeline, alongside breaking renames to section-label fields, schema route classes, and summarize_sections scoping.

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

01 LLM coreference resolution for fact merging NEW 93

The `AGG_LLM_COREFERENCE` node (on by default) adds one LLM call per document after merging to settle entity pairs that symbolic tiers cannot resolve; each verdict either merges two entities like a shared key or vetoes the merge. The number of pairs sent per call is capped with `AGG_LLM_COREFERENCE_MAX_PAIRS`, and results are tracked via counters `facts_coreference_*` and phase `facts_coreference`.

Cap the entity pairs sent per coreference LLM call to control cost on large documents.

```
$ AGG_LLM_COREFERENCE_MAX_PAIRS=20 ontocast process <input>
```

— Named node, env var, counters, and a runnable example. v0.9.0

02 Routing and labelling diagnostics in run manifest and CLI

NEW 70

Every content unit now records why it was routed in `route_reason`; the run manifest's `selection` block adds `route_histogram` (units and characters per route and reason) and, under `target_sections`, characters removed and how many were unlabeled. When `CHUNK_SECTION_CLASSIFIER=llm`, `selection.outline_labels` records outline windows sent and failed by error class, exposed via counter `chunk/outline_windows_failed`. The `ontocast sections` command prints each unit's route/reason and the outline-label stats.

— Several named manifest fields and a CLI command, little mechanism. v0.9.0

03 Embedding-based heading label inference IMPROVED 65

At the `heuristic` and `llm` labelling levels, an unnamed heading now takes its nearest content label by embedding (using the chunker's model) as a weak vote via

source `heading_embedding`, applied when that label is close and clearly ahead of the next candidate; numbered subsections inherit the inferred label weakly.

— Mechanism explained but no user-facing control or command. v0.9.0

04 **Section label strength replaces confidence score** BREAKING

60

Content units no longer carry `section_label_confidence`; it is replaced by `section_label_strength` (`strong` or `weak`) and `section_label_evidence`. The provenance graph field `ontocast:sectionLabelConfidence` is replaced by `ontocast:sectionLabelStrength`, and the run manifest's `selection` adds `label_strength_chars`. Any pipeline reading either old field name will break.

— Exact old/new field names with explicit break warning. v0.9.0

05 **Section schema route class replaces default_exclude** BREAKING

60

Section schemas must now mark each label's route class (`content`, `boilerplate`, or `citations`) in place of `default_exclude`; schemas still using `default_exclude` will need to be updated.

— Named schema values and migration instruction. v0.9.0

THINNER COVERAGE BELOW

06 **Table of contents and list recognition** NEW

55

Tables of contents and lists of figures or tables are now recognised from their content — lines or two-column table rows ending in ascending page numbers — and routed as non-content with reason `contents_list`.

— Clear detection heuristic but no actionable surface. v0.9.0

07 **summarize_sections scope change** BREAKING

45

`summarize_sections` no longer limits which sections are extracted; it now only controls which units are summarized, decoupling extraction scope from summarization scope.

— Named config key with before/after but no usage example. v0.9.0

└──▷ **BREAKING ON UPGRADE**

! Content units no longer carry `section_label_confidence`; it is replaced by `section_label_strength` and `section_label_evidence`. The provenance graph field `ontocast:sectionLabelConfidence` is replaced by `ontocast:sectionLabelStrength`. Any pipeline reading either name will break.

! Section schemas must now mark each label's route class (`content` , `boilerplate` , or `citations`) in place of `default_exclude` ; schemas using `default_exclude` will need to be updated.

! `summarize_sections` no longer limits extraction to the sections it selects; it now only controls which units are summarized.

WAS THIS USEFUL?  

knowledge-rag

Feedback ⓘ v5.0.0 NOTES ↗ RANK

Local RAG MCP server for Claude Code — hybrid search (semantic + BM25), cross-encoder reranking, 13 MCP tools, 20 format parsers. Zero external servers, zero API keys.

> knowledge-rag's v5.0.0 release adds precise match positioning to search snippets and introduces pagination to `list_documents`, alongside a breaking config fix for multilingual profile setups.

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

01 Pagination for list_documents BREAKING

90

`list_documents` now returns paginated results with `total`, `offset`, `limit`, and `count` fields instead of the full set, with a default page size of 100 and a maximum of 200 via the `limit` parameter. This is a breaking change: callers that expected all documents in one response must now page through results by incrementing `offset` by the returned `count` until `offset` reaches `total`.

Page through all indexed documents when you have more than 100 and need to retrieve the full list.

```
$ # First page
curl 'http://localhost:8000/list_documents?limit=200&offset=0'
# Next page — increment offset by the returned count until offset
>= total
curl 'http://localhost:8000/list_documents?limit=200&offset=200'
```

— Named fields, default/max limits, and a runnable paging example. v5.0.0

02 Migration fix for multilingual profile config BREAKING

75

Installs that enabled `profile: "multilingual"` using the old `config.example.yaml` must delete the empty `query_prefix` / `passage_prefix` lines from their config and run reindex_documents(force=True) to pick up the fix.

— Exact config keys and migration command given, but no explanation of root cause. v5.0.0

THINNER COVERAGE BELOW

03 Match-offset positioning in search snippets NEW

50

Search snippets now include a `match_offset` field that positions the snippet at the start of the matched text rather than at the beginning of the passage, making it easier to jump directly to the relevant text in long documents.

— Names the field and behaviour but no usage example. v5.0.0

└──▷ **BREAKING ON UPGRADE**

! `list_documents` now returns at most 100 documents by default (up to 200 with `limit`), along with `total`, `offset`, and `limit` fields. Callers that expected all documents in one response must now paginate by incrementing `offset` by `count` until `offset` reaches `total`.

! Installs that enabled `profile: "multilingual"` using the old `config.example.yaml` must delete the empty `query_prefix` / `passage_prefix` lines from their config and run reindex_documents(force=True).

WAS THIS USEFUL? 👍 👎

AWS Context Ontology Accelerator

Feedback ⓘ

v0.3.9 CODE ↗ RANK

Context Ontology Accelerator provides an open-source semantic layer that helps AI agents make consistent, explainable decisions using formal ontologies and knowledge graphs.

> Context Ontology Accelerator's v0.3.9 release adds a bounded document-reading tool for the agentic tier with new window-sizing and chunk-ordering controls, a SPARQL firewall that restricts LLM-generated queries to read-only forms with named-graph scoping, and two new agentic retriever tools for document search and ontology lookup.

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

01 **Document reading tool in agentic tier** NEW 90

Adds `ReadDocumentTool` (registered as `READ_DOCUMENT_TOOL`) to the agentic tier, letting the planner follow a retrieved chunk back to its source document and read the full text one bounded window at a time, using `source_key` (an S3 key) or `doc_id` as the reference, with reads scoped to the bucket and `s3Prefixes` of a DOCUMENTS source registered to the request namespace. New `DEFAULT_MAX_CHARS` and `DOCUMENT_CHUNK_PREFIX` constants and a `rank_source_documents` function in `document_tool` control the window size and chunk ordering used by this reader.

— Names tool, registration key, config params and scoping mechanism v0.3.9

02 **SPARQL firewall for agentic graph queries** NEW 85

Adds `SPARQLFirewall` to gate LLM-generated SPARQL queries before they reach a graph endpoint, enforcing read-only forms (SELECT, ASK, CONSTRUCT, DESCRIBE only — INSERT, DELETE, DROP, CREATE, LOAD, CLEAR, COPY, MOVE, ADD are rejected) and named-graph scoping via `configuration.dataNamedGraph`, with caller-written `FROM` / `FROM NAMED` clauses always rejected.

— Clear enforcement rules and config key, but no runnable example v0.3.9

THINNER COVERAGE BELOW

03 **New document and ontology lookup tools** NEW 40

Adds `FIND_DOCUMENTS_TOOL` and `ONTOLOGY_LOOKUP_TOOL` to the agentic retriever, joining the existing `NL_TO_SPARQL_TOOL` and `NL_TO_SQL_TOOL`.

— Names the tools but gives no mechanism or usage detail v0.3.9

WAS THIS USEFUL? 👍 👎

okf-memory OKF Agent Memory

Feedback ⓘ v0.8.0

NOTES ↗ RANK

Git-native persistent memory for AI coding agents. 2 with sub-300µs in-memory BM25 search, embedded MCP server, and progressive disclosure.

> OKF Agent Memory ships declarative bundle scope isolation to lock which memory layers agents can read, a two-stage lockfile-and-checksum mechanism to prevent concurrent write corruption, and automatic SSoT tool symlinks in `okf bootstrap`, while removing the deprecated `bundle` scope alias.

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

01 **Bundle-level scope isolation** NEW 95

Restrict which memory scopes agents can access per project by declaring a `scopes` list in `knowledge/index.md` frontmatter; the four canonical scopes are `project`, `vendor`, `user`, and `system`, and omitting `scopes` keeps all four active. Omitted scopes are blocked even when queried explicitly via `--scope`, `okf show <scope>:<id>`, `okf search --scope user`, `okf show user:secret`, or `okf://user/secret`, returning a descriptive error naming the allowed scopes. Bundle loading and `okf validate` fail immediately (fail-closed) when `scopes` is set to an empty list (`scopes: []`) or contains non-canonical scope names.

Lock a client repository to only its own knowledge and vendored packages, preventing agents from reading user- or system-level memory.

yaml

```yaml
# knowledge/index.md
---
okf_version: "0.2"
scopes: [project, vendor]
---
```

Prevent agents in a client or open-source repository from reading personal user or system notes by locking the bundle to only project and vendor scopes.

yaml

```yaml
# knowledge/index.md
okf_version: "0.2"
scopes: [project, vendor]
```

— Names exact config key, file, CLI flags and failure modes. v0.8.0

02 **Automatic SSoT symlinks in okf bootstrap** NEW 90

`okf bootstrap` now automatically creates SSoT tool symlinks (`CLAUDE.md`, `.cursorrules`, `.windsurfrules`, and `.github/copilot-instructions.md`) pointing to `AGENTS.md`; pass `--no-symlinks` to skip symlink creation, e.g. `okf bootstrap /path/to/project --name "My Project" --no-symlinks`.

Bootstrap a new project without creating SSoT symlinks when the repo already manages its own CLAUDE.md and .cursorrules.

```
$ okf bootstrap /path/to/project --name "My Project" --no-symlinks
```

— Exact flag and runnable bootstrap command given. v0.8.0

03 **Removal of `bundle` alias for `--scope`** BREAKING 90

The `bundle` alias for `--scope` is removed from `okf search` and the MCP `okf_search` tool. `--scope` now accepts only `all`, `project`, `vendor`, `user`, or `system`; passing `--scope bundle` or `"scope": "bundle"` returns an invalid-scope error. Replace `--scope bundle` with `--scope project`.

— Exact flag values and migration replacement given. v0.8.0

04 **Two-stage concurrent write protection** NEW 80

Local writes are now protected by a cross-process bundle lockfile (`.okf-write.lock`) that serializes `InitBundle`, `SaveConcept`, `UpdateConcept`, `RelateConcepts`, `UpdateParentIndex`, and `AppendLogEntry`, plus an optimistic concurrency check (`ErrConcurrentModification`) that rejects stale overwrites when a concept's on-disk SHA-256 digest changes between load and save. `okf update` (`okf_update`) and `okf relate` (`okf_relate`) now acquire the bundle write lock before loading and mutating the target concept, preventing concurrent field updates or link insertions from overwriting each other.

— Mechanism and function names given but no direct user command. v0.8.0

└──▷ **BREAKING ON UPGRADE**

! The `bundle` alias for `--scope` is removed from `okf search` and the MCP `okf_search` tool. `--scope` now accepts only `all`, `project`, `vendor`, `user`, or `system`; passing `--scope bundle` or `"scope": "bundle"` returns an invalid-scope error. Replace `--scope bundle` with `--scope project`.

WAS THIS USEFUL? 👍 👎

helix-db

Feedback ⓘ v3.6.0 NOTES ↗ RANK

HelixDB is an OLTP graph database with native vector and full-text search built in Rust on Object Storage.

> helix-db v3.6.0 adds a SQL query interface that translates to Cypher, new diagnostics tooling, expanded authentication sign-in methods, and more consistent secondary index builds under heavy write load.

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

01 **SQL query support via POST /v2/sql** NEW 85

Query the graph with SQL through `POST /v2/sql`, which translates SQL to Cypher under the hood. SQL explain, gRPC, embedded, and SDK support are also included.

Query a helix-db instance using SQL when you do not want to write Cypher or use the SDK query DSL.

```
$ curl -X POST http://localhost:6969/v2/sql \
  -H 'content-type: application/json' \
  -d '{"query": "SELECT * FROM User"}'
```

— Names endpoint and translation mechanism with a runnable example v3.6.0

02 **Instance diagnostics via helix doctor and /v2/diagnostics** NEW 75

Run `helix doctor` from the CLI or call `GET /v2/diagnostics` to get instance diagnostics without connecting to the database manually.

Check the health and configuration of a running helix-db instance from the CLI.

```
$ helix doctor
```

Inspect instance diagnostics programmatically, e.g. from a monitoring script or health-check endpoint.

```
$ curl http://localhost:6969/v2/diagnostics
```

— Exact CLI command and endpoint with curl examples v3.6.0

THINNER COVERAGE BELOW

03 **Secondary index consistency under sustained writes** IMPROVED

35

Secondary index builds are now activated under sustained writes, enabling index consistency in write-heavy workloads.

— Describes behaviour change but no config or command to act on v3.6.0

04 **Expanded sign-in methods in helix auth login** IMPROVED

30

`helix auth login` now supports every sign-in method, expanding the methods previously available.

— Names the command but not which methods were added v3.6.0

WAS THIS USEFUL? 👍 👎

Docling

Feedback ⓘ v2.148.0 NOTES ↗ RANK

Docling parses PDFs, Office files, and HTML into structured documents with layout, tables, and reading order preserved for downstream RAG pipelines.

> Docling's latest releases add a new iCalendar backend for parsing calendar files, bring surya v2.5 support for layout and OCR, and introduce superscript/subscript detection in PDFs.

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

01 **iCalendar file parsing backend** NEW 80

Docling adds a new backend for calendar files in `.ics`, `.ical`, `.icalendar`, and `.ifb` formats, extracting events, tasks, journal entries, free/busy times, attendees, reminders, and attachment names.

Parse a calendar file into Markdown to extract meeting and event text from an .ics file.

python

```python
from docling.document_converter import DocumentConverter

converter = DocumentConverter()
result = converter.convert("team_schedule.ics")
print(result.document.export_to_markdown())
```

– Names formats and extracted fields, with a working code example. v2.148.0

THINNER COVERAGE BELOW

02 **Superscript and subscript detection in PDFs** NEW 45

Docling's PDF backend now marks superscript and subscript text by inferring it from glyph size and baseline position.

– Explains detection mechanism but gives no usage example. v2.148.0

03 **Surya v2.5 OCR and layout support** IMPROVED 30

Docling now supports surya v2.5 for improved layout and OCR analysis.

– Names the engine version but no mechanism or usage detail. v2.148.0

WAS THIS USEFUL? 👍 👎

rabitq-library

Feedback ⓘ v0.8.3 NOTES ↗ RANK

An official lightweight library for the RaBitQ algorithm and its applications in vector search.

> rabitq-library exposes quantization primitives directly to Python via a new provisional module.

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

01 **Provisional Python quantization module** NEW 50

Adds the provisional `rabitqlib.quant` module, giving Python access to quantization primitives directly.

— Names the module but no usage details or API surface. v0.8.3

WAS THIS USEFUL? 👍 👎

fastembed-rs

Feedback ⓘ v7.6.0 NOTES ↗ RANK

Rust library for generating vector embeddings and reranking locally!

> fastembed-rs v7.6.0 adds automatic ONNX Runtime device selection, letting the library pick CPU, GPU, or another available accelerator.

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

01 **Automatic device selection via InitOptions** NEW 55

ONNX Runtime can now automatically select the execution device (CPU, GPU, or other available accelerator) by configuring it through `InitOptions`.

— Names config surface but no code example or defaults given. v7.6.0

WAS THIS USEFUL? 👍 👎

infino

Feedback ⓘ v0.11.25 NOTES ↗ RANK

Embedded retrieval library built on Parquet. Fast, efficient, and scalable.

> infino added persistent on-disk vector index state and extended its catalog with `information_schema` support for schema metadata queries.

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

01 **information_schema support in catalog** NEW 40

infino's catalog now supports `information_schema`, making schema metadata queryable alongside regular tables.

— Names the catalog surface but no query examples given v0.11.25

02 **Persistent vector index state on disk** NEW 20

infino now persists hidden vector index state on disk.

— Only a bare mention with no mechanism or usage detail v0.11.25

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

Comet Opik

Feedback ⓘ 2.2.143 NOTES ↗ RANK

Opik traces, evaluates, and monitors LLM applications and agentic workflows, with datasets, an LLM-as-judge metric library, and production dashboards.

> Opik shipped service account token authentication, guaranteed persistence for online-evaluation monitoring traces, Decision rule scoring via OpenAI's Decisions API, and the ability to jump from a trace view straight into the playground for prompt iteration.

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

01 **Decision API scoring with gpt-6-luna** NEW 60

Opik can now score online-evaluation Decision rules using OpenAI's Decisions API with the gpt-6-luna model.

— Names the API and model, no config key given 2.2.143

THINNER COVERAGE BELOW

02 **Open LLM spans in playground from trace view** NEW 50

Users can open any LLM span directly in the playground from the trace view for quick prompt iteration.

— Clear UI navigation path but no exact steps 2.2.142

03 **Service account token authentication** NEW 45

Opik now supports service account token authentication, returning the plain-text token in API replies.

— Names the mechanism but no endpoint or flag given 2.2.143

04 **Persistent online-evaluation monitoring traces** IMPROVED 35

Opik now always persists online-evaluation monitoring traces so they are never dropped.

— States the behaviour change but no config exposed 2.2.143

05 **Filter experiments by evaluation method** NEW 30

Experiments can now be filtered by evaluation method in the backend.

— Bare one-line addition, no UI path or endpoint named 2.2.143

06 **Organization context for backend requests** NEW 25

Opik now adds an organization context with a default organization to backend requests.

— Bare description with no mechanism detail 2.2.143

WAS THIS USEFUL? 👍 👎

OpenCompass

Feedback ⓘ 0.5.5 NOTES ↗ RANK

OpenCompass evaluates language and multimodal models across datasets, inference backends, and benchmark metrics.

> OpenCompass expanded its benchmark catalog with 19 new configurations spanning physics, reasoning, long-context, alignment, hallucination, maritime, biomedical and translation evaluation, added multimodal datasets through VLMEvalKit, improved TurboMind API diagnostics, and dropped Python 3.8 support.

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

01 **Retry and verbose options for TurboMind API** IMPROVED 60

Added retry and verbose options to the TurboMind API for `get_ppl` and `get_ll`.

– Names exact API methods but not the flags themselves 0.5.5

THINNER COVERAGE BELOW

02 **New benchmark configurations across domains** NEW 55

Added benchmark configurations for `CritPt`, `General365`, `AA-OmniScience` (v1.0 and v1.1), `AA-LCR v1.1`, `HLE-Diamond`, `arena-hard-v2`, `ReasonZoo`, `LG-Bench`, `CHHallu-Src v1`, `GrandPhysics`, `PromptCBLUE Life Science`, `MaritimeBench`, `WMT19`, `LogiQA`, and an ethics evaluation configuration, part of a batch of 19 new benchmark configurations spanning physics, reasoning, long-context, alignment, hallucination, maritime, biomedical and translation evaluation.

– Names every benchmark config but no usage mechanism given 0.5.5

03 **Multimodal datasets via VLMEvalKit** NEW 45

Added `ChartQAPro`, `MicroVQA`, `SFE`, and `SciDocBench` multimodal datasets, integrated through VLMEvalKit.

– Names datasets and integration point but no setup detail 0.5.5

04 **Python 3.8 support dropped** BREAKING 45

Python 3.8 is no longer supported; installations running Python 3.8 will break on upgrade.

– Clear migration note but no replacement guidance given 0.5.5

└──▷ **BREAKING ON UPGRADE**

! Python 3.8 is no longer supported; installations running Python 3.8 will break on upgrade.

WAS THIS USEFUL? 👍 👎

Arize Phoenix

Feedback ⓘ @arizeai/phoenix-cli@1.20.0 NOTES ↗

RANK

Arize Phoenix is an open-source platform that monitors LLM application traces and evaluates their outputs.

> Arize Phoenix's CLI gained a command to list a prompt's versions with pagination and flexible output formats.

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

01 **CLI command to list prompt versions** NEW 83

The new `px prompt-versions list <prompt-identifier>` command lists all versions of a named prompt, supporting cursor pagination, optional limits, and `pretty`, `JSON`, and `raw` output formats, with creation dates shown when the server supports them.

— Names exact command, flags, and output formats; runnable as-is.

@arizeai/phoenix-cli@1.20.0

WAS THIS USEFUL? 👍 👎

Langfuse

Feedback ⓘ v4.76.0 NOTES ↗ RANK

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

> Langfuse's v4.76.0 release adds Topics support to the public API and MCP, ships the AI gateway in self-hosted Docker Compose, and introduces audit logging for trace reads and data exports, alongside several UI and alerting improvements.

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

01 **Audit-log events for trace reads and exports** NEW 50

Langfuse now records read-access audit-log events for tRPC trace, session, and score reads, and separately records exports and downloads as audit-log access events.

— Scope named but no API or config surface specified. v4.76.0

02 **Topics support in public API and MCP** NEW 45

Langfuse now exposes Topics summaries, assignments, and rules through the public API, the MCP server, and the in-app agent.

— Names the surfaces but no endpoint paths or schemas given. v4.76.0

03 **Session list view and trace peek session tab** NEW 45

The sessions UI gained a new chronological session list view, and the trace peek view now includes a dedicated session tab.

— UI additions described with location but no deeper mechanism. v4.76.0

04 **PagerDuty monitor alerts with custom details** NEW 40

Monitor alerts can now be sent to PagerDuty with custom details and matching Slack links.

— Feature named but no setup steps or config keys given. v4.76.0

05 **AI gateway added to self-hosted Docker Compose** NEW 35

The AI gateway is now shipped as part of the self-hosted Docker Compose deployment setup.

— Clear deployment starting point but no config detail. v4.76.0

06 **Searchable org, project, and account switcher** NEW 35

The web UI adds a searchable switcher for navigating between organizations, projects, and accounts.

— Simple UI addition with minimal supporting detail. v4.76.0

WAS THIS USEFUL? 👍 👎

overmind

Feedback ⓘ v0.1.29 NOTES ↗

The platform for continuously improving AI agents.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.1.29 https://github.com/overmind-core/overmind.git
# already have the repo? check out this version:
$ git checkout v0.1.29
```

NEW SUPPORTED PLATFORM LFM3-1.7B LoRA training and SGLang serving

WAS THIS USEFUL? 👍 👎

Traceloop OpenLLMetry

Feedback ⓘ 0.63.0 NOTES ↗ RANK

OpenLLMetry instruments LLM and vector-database calls as OpenTelemetry spans so traces land in existing observability backends rather than a new one.

> OpenLLMetry added a `trace_content` parameter to control whether prompt and response content is captured in traces, letting teams avoid storing sensitive data in their observability backend.

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

01 **Trace content toggle in Traceloop.init()** NEW 75

A new `trace_content` parameter can be passed to Traceloop.init() to control whether LLM prompt and response content is captured in spans, e.g. Traceloop.init(trace_content=False) to disable capture and avoid storing sensitive data in the observability backend.

Disable prompt/response content capture in production to avoid storing sensitive data in your observability backend.

python

```python
from traceloop.sdk import Traceloop

Traceloop.init(trace_content=False)
```

– Names exact init parameter and usage; lacks broader scope detail 0.63.0

WAS THIS USEFUL? 👍 👎

PromptLayer

Feedback ⓘ November 05, 2026 NOTES ↗ RANK

PromptLayer is a platform for managing, evaluating, and observing prompts and LLM applications.

> PromptLayer now lets users import Smart Table rows directly as input sets in the prompt editor, streamlining batch testing with real data.

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

01 **Import Smart Table rows as input sets** NEW 84

Using the Table rows option in the Open Library menu, users can import rows from a Smart Table directly as input sets, with each selected row becoming its own named input set. Columns matching a prompt's variable names are imported automatically, and when none match, all columns are imported; structured values such as JSON are imported as readable, formatted text rather than raw strings. Row previews show each row's first three columns to help identify the right row, and selection is limited to the free input set slots, with a maximum of five sets.

Quickly populate a prompt's input sets with real data rows from a Smart Table for batch testing.

> 📍 In the prompt editor, open the Open Library menu, select Table rows, choose your sheet, review the row previews showing each row's first three columns, select up to five rows, and confirm — each row becomes its own input set named after the table and row number.

— Names exact menu path, limits, and behavior; missing only naming convention detail.

November 05, 2026

WAS THIS USEFUL? 👍 👎

Braintrust

Feedback ⓘ updated · Nov 6, 2026 NOTES ↗

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

Braintrust adds trace continuation via `--parent` in `bt trace run`, MCP job-run listing, a thread view for human review, and Patterns detection of user-noticed problems.

Continue a trace from an existing span by passing `--parent` to `bt trace run` with the value returned by an SDK span's export(), attaching a coding-agent session below that span instead of starting a new trace.

Call a Braintrust MCP server tool with only its own arguments — `bt mcp run` now makes the telemetry argument optional.

List the past runs of a scheduled Loop automation with `list_scheduled_loop_job_runs` on the Braintrust MCP server; each run includes a link that opens it in Braintrust (requires data plane v2.22.0+).

In human review, read traces in the new thread view (public preview) and collapse the scores panel to give the trace more room.

The search dropdown's Recent section now includes structured filters alongside SQL filters, so a recent filter can be reapplied without rebuilding it.

+1 more

UPDATED INTEGRATION Anthropic Java SDK 3.x tracing support

NEW INTEGRATION Temporal activity span recording in Python SDK

NEW INTEGRATION OpenAI Responses API span recording in Python SDK

NEW INTEGRATION Cursor subagent and OpenCode model call span recording

UPDATED INTEGRATION Gemini thinking token recording in Java SDK

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI/LLM Security

cplt

Feedback ⓘ 2026.11.05-132144-6424a3a 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 adds a way to share a linked repository with teammates via version-controlled configuration.

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

01 **Propose flag for shared repository links** NEW 70

The `link` command gains a `--propose` flag that records the link in `.cplt.toml`, enabling version-controlled distribution of a linked repository to teammates.

— Names exact flag and config file, lacks broader mechanism detail

2026.11.05-132144-6424a3a

WAS THIS USEFUL? 👍 👎

containarium

Feedback ⓘ v0.131.0 NOTES ↗ RANK

Open-source agent runtime — SSH-native isolation, eBPF egress policy, Kubernetes + LXC backends, GPU passthrough, MCP-native CLI

> Containarium's v0.131.0 release adds large-file streaming into container boxes and a new MCP tool for in-box uploads, plus a fix to REST error handling and visibility into installed coding agents.

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

01 **put_file MCP tool for in-box uploads** NEW 75

Adds a `put_file` MCP tool to `agent-box`, backed by `code put-file`'s client function, for uploading files into a running container box, with the per-call cap documented. It can be wired into coding agents like Claude Code by configuring an SSH-based MCP server pointing at `agent-box`.

Upload a large file (over 2 MiB) into a running container box via the new `put_file` MCP tool wired through `agent-box` over SSH.

```
$ ssh user@your-box agent-box
```

Wire `agent-box` into Claude Code so the `put_file` tool is available for in-box file operations from your AI coding agent.

```json
{
  "mcpServers": {
    "containarium-box": {
      "command": "ssh",
      "args": ["user@your-box", "agent-box"]
    }
  }
}
```

— Named tool and config example, but exact cap value missing v0.131.0

THINNER COVERAGE BELOW

02 **Streamed large file writes via WriteBoxFileStream** IMPROVED 52

Adds `WriteBoxFileStream` to stream files larger than 2 MiB into a container box, alongside a fix for REST box-file error decoding.

— Names the API and size threshold but no usage path v0.131.0

03 **Installed agent recorded after code install** IMPROVED

40

Records the installed coding agent after `code install` verifies, making post-install state visible.

— States behaviour but no detail on where state is surfaced v0.131.0

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Anthropic

Feedback ⓘ updated · Nov 6, 2026 NOTES ↗ RANK

The Claude Developer Platform is the API for building applications on Anthropic's Claude models that generate, analyze, and automate work.

> Anthropic strengthened admin controls for the Claude Developer Platform with organization-wide IP allowlisting and a remote sign-out capability for Enterprise accounts.

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

01 **IP allowlisting for API requests** NEW 70

Organizations can set up an IP allowlist in the Claude Console to refuse API requests from addresses not on the list, limiting where a leaked API key can be used. It supports IPv4 and IPv6 addresses or CIDR ranges, with up to 1,000 entries per organization.

— Names mechanism, formats and the 1,000 limit; only UI path given snapshot-20261106

THINNER COVERAGE BELOW

02 **Remote sign-out via Admin API** NEW 45

The Admin API can now sign a Claude Enterprise organization member out of Claude everywhere without removing them from the organization, allowing the member to sign back in afterward.

— Explains behavior but no endpoint path or parameters given snapshot-20261106

WAS THIS USEFUL? 👍 👎

Google Gemini Omni 1.1 Flash

Feedback ⓘ

released · Nov 6, 2026 MODEL ↗

Gemini Omni 1.1 Flash from Google was released on 4 Nov 2026. 131K-token context; $1.50 in / $9.00 out per million tokens (OpenRouter).

WAS THIS USEFUL? 👍 👎

Google Gemini Omni 1.2 Flash

Feedback ⓘ

released · Nov 6, 2026 MODEL ↗ NOTES ↗

Gemini Omni 1.2 Flash from Google was released on 5 Nov 2026. 131K-token context; $1.50 in / $9.00 out per million tokens (OpenRouter).

Succeeds Gemini Omni 1.1 Flash.

Gemini API adds Gemini Omni 1.2 Flash GA with up to 15-second video generation, conversational editing, and 30% lower output pricing

Gemini Omni 1.2 Flash is now generally available via model ID `gemini-omni-1.2-flash`, supporting up to 15-second video generation, conversational editing (3–10 seconds for first and last frame interpolation, 3–10 seconds per-turn video extension up to 30 seconds total), improved character and subject consistency across up to 6 reference images, 30% lower video output pricing ($12.00 per 1M tokens), and doubled paid-tier rate limits.

└─▷ **BREAKING ON UPGRADE**

! Effective December 3, 2026, API requests to `gemini-omni-1.1-flash` will automatically route to `gemini-omni-1.2-flash` and be billed at Gemini Omni 1.2 Flash rates.

WAS THIS USEFUL? 👍 👎

◆ **Open-Weight Models**

OmniJev OneJev1.5g-31B

Feedback ⓘ released · Nov 6, 2026

MODEL ↗

OneJev1.5g-31B from OmniJev was released on 5 Nov 2026. Open weights on Hugging Face.

WAS THIS USEFUL? 👍 👎

Z.ai GLM 5.3 Feedback ⓘ price cut · Nov 5, 2026 MODEL ↗

GLM 5.3 is a reasoning model from Z.ai.

Price cut on OpenRouter (z-ai/glm-5.3): $0.08 → $0.0699 in, $3.60 → $1.40 out per million tokens (-61%), first seen 5 Nov 2026.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · November 6, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue