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

THE AI TOOLCHAIN

NO. 082

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED NOVEMBER 3, 2026 · EVERY DAY

EDITIONS **tail** | grep | head | ls | diff | uniq | perf

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 405 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 101 of 105 tools

▾ // THE BRIEF

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

Three unrelated things stand out today: Anthropic stores conversation state server-side so long agent loops stop resending history; vLLM adds more speculative decoding paths to cut inference latency; brightdata-mcp parallelizes search and page fetches so data-gathering agents wait on batches, not pages.

OPERATE

Stop resending the whole conversation on every agent turn

Anthropic's server-side message threads let clients send only the new messages for a continuing conversation, which removes repeated large payload uploads and should improve time to first token in long loops. PydanticAI's persistent ConversationStore with conflict detection gives app code a safer place to keep that state instead of hand-rolling replay and overwrite handling.

Anthropic · PydanticAI

DEPLOY

Serve the same model faster on the same hardware

vLLM added several speculative decoding throughput techniques, where a cheaper draft path proposes tokens and the main model verifies them to reduce end-to-end latency. SGLang added low-latency decode plugins and broader speculative decoding and caching gains, giving serving teams more options before adding GPUs.

vLLM · SGLang

BUILD

Let agents collect web evidence in fewer serial steps

brightdata-mcp added parallel search and batch page-fetch tools with configurable formats and timeouts, so an agent can fan out research work instead of issuing one blocking browser-style request at a time. Recency filtering on the dedicated SERP API zone also makes freshness an explicit query control rather than a post-filtering chore.

brightdata-mcp

GOVERN

Expose internal AI services without making every route public

SAM added policy-gated HTTP Ingress for a2a, mcp and inference services, so mesh services can be reached from outside through an explicit policy boundary. hayhooks added production controls to disable

> runtime pipeline deploy and undeploy over HTTP, CLI and MCP, preventing a live server from being reshaped after startup.
>
> SAM

Does **GPT-5.5** read better? 👍 👎

APPLICATION SECURITY

◆ SAST / SCA / Code Security

opentaint

Feedback ⓘ go-server/2026.11.02.54ad708 NOTES ↗ RANK

The open source taint analysis engine for the AI era. A formal dataflow analysis tool you can customize and self-host, built so AI agents drive your application security analysis without burning tokens on every scan.

> opentaint's analyzer gained built-in Go models for taint analysis, extending coverage beyond Java, alongside a breaking terminology change renaming approximations to models throughout the tool.

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

01 **Renamed "approximations" to "models"** BREAKING 50

The concept previously called approximations is renamed to models throughout the tool; any configuration, rules, or scripts referencing the old term approximations will need to be updated.

– Clear migration note but no exact config keys listed analyzer/2026.11.02.54ad708

02 **Go models added to taint analyzer** NEW 32

opentaint now includes built-in Go models for taint analysis, extending model coverage beyond Java, and the analyzer gained general model support.

– States what changed but no mechanism or config detail analyzer/2026.11.02.54ad708 go-server/2026.11.02.54ad708

└──▷ BREAKING ON UPGRADE

! The concept previously called approximations is renamed to models throughout the tool; any configuration, rules, or scripts referencing the old term approximations will need to be updated.

WAS THIS USEFUL? 👍 👎

BUILD

◆ AI Coding Agents

yetone magpie

Feedback ⓘ v0.1.2826 CODE ↗ RANK

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

> Magpie shipped a dense set of gateway and UI improvements this window, headlined by a memory-aware body-size limit that lets large Codex sessions complete, per-key USD spend caps, and ChatGPT backend-native web search that costs no model tokens.

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

01 **Context-window archive fallback for expired requests** IMPROVED

82

Context-window cards now read a request's text and reply from the request archive via `/api/context/<what>` when the in-memory copy has expired (magpie keeps only the last 10 requests' bodies in memory), with secrets masked the same way as in memory. A 404 from that endpoint now returns a JSON `{archive: why}` field explaining the gap — `off` (archive disabled or route missing), `missing` (not yet saved or removed), `cut` (only the start was kept), or `unread` (could not be read or masked).

— Named endpoint and enumerated error reasons v0.1.2808

02 **Library skill file read and edit in UI** NEW

80

Read a library skill's files in the UI via the `ReadSkillFile` function, which returns the file content, path, and a list of all text files for the skill (capped at `skillFilesMax` = 200 files, `skillFileMax` = 1 MB per file), with `SKILL.md` listed first. Skills can also be edited directly in the library page — saves go to the library's folder so every agent linked to that skill receives the update without manual sync.

— Named function, limits, and propagation mechanism v0.1.2825

03 **Memory-aware body-size limit for Codex requests** IMPROVED

77

The Codex backend path now reads request bodies up to a sixteenth of the machine's total RAM via codexBodyLimit(), with a floor of `defaultBodyLimit` (512 MB) and a ceiling of `codexBodyMost` (2 GB), letting long image-heavy Codex sessions that previously received a 413 complete. Requests that name a magpie model are still capped at the standard per-route limit via bodyLimitOf(), and now return a descriptive 413 stating both the magpie-model limit and the higher Codex-native limit for comparison.

— Full mechanism and named functions, but behavior is automatic v0.1.2823

04 **Per-key USD spend caps** NEW 75

Set a USD spend cap on any API key — daily (`SpendDay`), monthly (`SpendMonth`), or lifetime (`SpendTotal`) — via `SetKeySpendCap` ; once the cap is reached, magpie routes to the provider's other keys or answers 429 if none remain, and the cap resets automatically when the period renews.

— Named function and fields, but no CLI/UI path shown v0.1.2819

05 **ChatGPT backend-native web search** NEW 75

Use `@search` as the model identifier in Settings' Searcher field (formatted `<provider>/@search`) to route web searches through the ChatGPT backend's own search path (`/alpha/search`) instead of a model turn, consuming no model tokens. A new `Search` field on `SearcherChoice` indicates when a provider searches via the ChatGPT backend rather than a model.

— Exact setting and format a reader can type today v0.1.2821

06 **Selective header forwarding to providers** NEW 70

Providers now support a `PassHeaders` list that names which incoming agent request headers (such as `X-Tenant` or `X-Trace`) are forwarded to the upstream vendor, so per-client or per-request values that a fixed `Headers` block cannot express can reach the provider, while magpie's own auth and connection headers are held back.

— Names the config field and example headers v0.1.2815

07 **Codex restart control on account switch** NEW 65

Control whether magpie restarts the Codex app when it switches Codex accounts on its own by setting `RestartApp` in `providers.json` via the new `SetRestartApp` function, or leave the app running on the old account until the user restarts it manually.

— Names config key, file and function to set it v0.1.2826

08 **Zhipu/Z.ai balance support** NEW 60

magpie now reads account balances for Zhipu/Z.ai pay-as-you-go keys via the `zhipuBalanceSource` integration, covering `open.bigmodel.cn` , `www.bigmodel.cn` , and `api.z.ai` .

— Named integration and domains, automatic feature v0.1.2824

09 **Claude Code statusline toggle** NEW 60

A new toggle on the Agents page enables the `magpie statusline` command inside Claude Code's status bar. magpie protects existing custom status lines (such as claude-hud or a user script) by hiding the toggle and refusing to overwrite `settings.json` when a non-magpie `statusLine` is already present.

— Names UI location, command, and settings file guarded v0.1.2822

THINNER COVERAGE BELOW

10 **DeepSeek thinking-block stripping for Claude switches**

IMPROVED 55

The gateway now filters out thinking blocks signed by DeepSeek (UUID-format signatures) before sending a request to Claude, preventing Claude's 400 'Invalid `signature` in `thinking` block' error when a routing group switches from DeepSeek to Claude mid-conversation.

— Explains exact error fixed but behavior is automatic v0.1.2806

11 **Rate-limit rest status on Providers page** NEW 50

The Providers page now shows rate-limit rest status for subscription accounts and their individual models on each account row, including the model name, reset time, and the vendor's original error message.

— UI location given but no deeper mechanism v0.1.2814

12 **SiliconFlow balance queries retired** BREAKING 45

SiliconFlow balance queries are retired (`siliconFlowRetired`); magpie no longer queries SiliconFlow's balance API and instead surfaces a clear message directing users to the SiliconFlow console when the query is unavailable.

— Clear behavior change but no migration steps beyond redirect v0.1.2824

└──▷ **BREAKING ON UPGRADE**

! SiliconFlow balance queries are retired (`siliconFlowRetired`); magpie no longer queries SiliconFlow's balance API and will instead show a console redirect message.

WAS THIS USEFUL?  

devoxxgenieideaplugin

Feedback ⓘ v1.16.0 NOTES ↗

RANK

DevoxxGenie is an agentic plugin for IntelliJ IDEA that uses local LLM's (Ollama, LMStudio, GPT4All, Jan and Llama.cpp) and Cloud based LLMs to help review, test, explain your project code. Latest version now also supports Spec Driven Development with CLI Runners.

> DevoxxGenie's headline addition this window is an experimental `run_script` agent mode that lets the model batch multiple tool calls into one sandboxed JavaScript snippet, paired with a fix that stops project context from being re-sent on every message.

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

01 Experimental `run_script` agent code mode NEW 90

Enabled via Settings → DevoxxGenie → Agent → Built-in Tools → run_script, this lets the model write a short async JavaScript snippet that calls other tools as await tools.read_file({ path }), loops over results, and returns only what it prints with text() or `console.log` — cutting round trips and keeping chat memory lean. A new Max tool calls per script setting (default 100) governs script-level calls separately from the existing Max tool calls per prompt limit, with output capped at 20,000 characters, memory at 128 MiB, and a 30-second JavaScript time budget. Scripts execute in a fresh QuickJS-over-WebAssembly engine via `quickjs4j` with no Node.js, `require`, timers, file system, or network access, and every inner tool call still goes through the same approval dialogs and activity rows as a direct call. The full `run_script` tool description is editable like other built-in tools, and `ALL_TOOLS` exposes the complete callable-tool catalog as TypeScript-style signatures at runtime.

Turn on `run_script` so the agent can read multiple files and filter results in one round trip instead of one tool call per file.

📍 In the IDE, go to Settings → DevoxxGenie → Agent → Built-in Tools and enable `run_script`. Then optionally raise Max tool calls per script above the default of 100 for larger tasks.

— Names exact settings, limits, engine, and functions; only lacks a single runnable command.

v1.16.0

THINNER COVERAGE BELOW

02 Single-shot project context in system prompt IMPROVED 55

Project context added via Add Project is now sent once in a single `<ProjectFilesContext>` section of the system prompt and updated in place by

`prepareMemory`, rather than being copied into every user message — preventing context windows from filling with stale project snapshots.

— Internal mechanism named but no user-facing action to take. v1.16.0

WAS THIS USEFUL? 👍 👎

planning-with-files

Feedback ⓘ v3.28.0 NOTES ↗ RANK

planning-with-files is a persistent file-based planning system for AI coding agents with crash-proof markdown plans, session recovery, and context-rot protection.

> planning-with-files adds opt-in archiving for named plans, letting users move inactive plans out of the way while keeping hook behavior sane when multiple plans exist.

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

01 Archive and restore for named plans NEW

90

New `--archive`, `--restore`, and `--list --archived` flags let users move a named plan and its sidecars to `.planning/.archive/<name>/` (e.g. `sprint-42`), restore it later, and list archived plans; the JSON inventory now reports an `archived_count` field.

Dry-run an archive of a named plan to confirm the destination path before committing the move.

```
$ set-active-plan.sh --archive sprint-42
```

Apply the archive after previewing, moving the plan and its sidecars to `.planning/.archive/sprint-42/`.

```
$ set-active-plan.sh --archive sprint-42 --yes
```

Verify which plans are archived and check the `archived_count` field in the JSON inventory.

```
$ set-active-plan.sh --list --archived
```

— Names exact flags, destination path, and inventory field with runnable commands.

v3.28.0

THINNER COVERAGE BELOW

02 Hook behavior for multiple unselected plans IMPROVED

45

When a project holds several named plans and no `PLAN_ID` is set, hooks now inject nothing instead of guessing, and the multiple-plans notice names the archive command

as a way to resolve the ambiguity.

— Describes behavior change but gives no concrete command to act on it. v3.28.0

WAS THIS USEFUL? 👍 👎

Anthropic Claude Code

Feedback ⓘ v2.1.317 NOTES ↗

CODE ↗ RANK

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

> Claude Code v2.1.317 adds a gateway cost-report admin API with per-IdP-group breakdown, an environment variable to control nested session identity, stricter Agent SDK hook failure handling, richer OTel billing/routing attributes, and several smaller safety and scheduling notices.

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

01 **Nested session identity environment variable** NEW 90

Control nested session behavior on self-hosted runners by setting `CLAUDE_CODE_DISABLE_NESTED_SESSION_ID=1` so `claude -p` run inside a session uses that session's ID and transcript instead of creating its own.

Prevent a `claude -p` subprocess on a self-hosted runner from forking the parent session's transcript into a separate conversation.

```
$ CLAUDE_CODE_DISABLE_NESTED_SESSION_ID=1 claude -p 'Run the test suite and summarize failures'
```

— Exact env var with a runnable example command v2.1.317

02 **Cost-report API with per-IdP-group breakdown** NEW 85

Read past spend from the Claude apps gateway via the new `GET /v1/organizations/analytics/cost` admin API endpoint, broken down by total, per developer, per day, or per IdP group over a date range. A daily history of each developer's IdP groups is now recorded on Claude apps gateways (requires `admin:` scope) and retained for `admin.identity_retention_days`, which powers the per-IdP-group breakdown in the cost report.

— Named endpoint and config key but no sample call shown v2.1.317

03 **Agent SDK hook failure blocking** IMPROVED 70

Agent SDK hook callbacks can now block an action on failure by setting `onFailure: "block"`; previously a failing callback let the action proceed anyway.

— Named config field but no code sample given v2.1.317

04 **OTel billing and routing attributes** NEW 65

Claude Code now tracks richer billing and routing context in OTel metrics, events, and spans via the new `anthropic.plan`, `anthropic.seat_type`, `anthropic.provider`, and `anthropic.custom_base_url` attributes.

— Named attributes but no setup instructions v2.1.317

THINNER COVERAGE BELOW

05 **Unrequested connector warning in /schedule** IMPROVED

55

`/schedule` now warns when a new cloud scheduled task receives connectors that were not requested, including instructions on how to remove them.

— Names the command but not the removal steps v2.1.317

06 **Link card for MCP server URL requests** NEW

45

Cloud sessions now show a link card when an MCP server asks you to visit a URL, letting you open or decline the request instead of it going unseen.

— UI behavior described without exact navigation path v2.1.317

07 **Confirmation before sending account email to network** IMPROVED

40

Claude Code now prompts for confirmation before a web or shell tool call sends your account email address to the network.

— No detail on trigger conditions or mechanism v2.1.317

WAS THIS USEFUL?  

wrongstack

Feedback ⓘ v1.4.0 NOTES ↗ RANK

An AI coding agent that reads your code, edits files, runs commands, and reasons through bugs — across a terminal REPL, a full-screen TUI, and a browser UI, while you keep your hand on every permission.

> WrongStack's SAGE memory subsystem gains a WrongMemory service backend, letting durable memory be handed off to an external service instead of staying purely local.

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

01 **WrongMemory service backend for SAGE** NEW 40

SAGE now supports a `WrongMemory` service backend with store handoff, enabling durable memory to be managed by an external WrongMemory service rather than only locally.

— Names the backend but no setup steps or config keys given. v1.4.0

WAS THIS USEFUL? 👍 👎

The Open Engine Zeroshot

Feedback ⓘ v11.0.0 NOTES ↗

RANK

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

> Zeroshot's two releases separate Git delivery confirmation from graph output, add a Microsoft Foundry provider for Codex and Claude, and expand per-node execution labeling and harness configuration across all six adapters.

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

01 Microsoft Foundry provider for Codex and Claude NEW

90

Codex and Claude runtimes can now be pointed at a Microsoft Foundry deployment by setting `provider: "foundry"` in the runtime config, authenticated via `AZURE_FOUNDRY_API_KEY` with either `AZURE_FOUNDRY_RESOURCE` or `AZURE_FOUNDRY_BASE_URL`.

Route a Codex node to a model deployed in your Microsoft Foundry resource instead of OpenAI directly.

json

```
{
  "harness": "codex",
  "provider": "foundry",
  "model": "my-foundry-deployment-name",
  "effort": "high"
}
```

— Exact config keys, env vars, and runnable example v10.25.0

02 Delivery receipt field decoupled from graph output BREAKING

88

A new optional `status.metadata.deliveryReceipt` field exposes Git delivery evidence independently of graph output, so a successful native Git delivery is no longer masked when a custom graph returns a summary or `null`; it is also surfaced in the Python SDK as `RunResult.delivery_receipt`. Delivery confirmation is now sourced from native execution history rather than graph output, so runs that previously resolved as `delivery_unconfirmed` may resolve differently, and any code inspecting graph output for delivery confirmation must switch to `status.metadata.deliveryReceipt`.

— Names exact fields, mechanism, and required migration v11.0.0

THINNER COVERAGE BELOW

03 **Execution labels across all harness adapters** NEW 55

Zeroshot now exposes seven `ZEROSHOT_*` execution labels across all six harness adapters, including reducer attempts and the innermost loop iteration, preserved through corrections, replay, capsule requests, and auxiliary launches.

— Counts and scopes labels but doesn't name them v10.25.0

04 **Per-node native harness configuration** NEW 45

Per-node native harness configuration is now available across all six adapters, including Claude verifier tool selection, with overrides preserved through profiles, corrections, and session reuse.

— Describes scope but no config key or flag named v10.25.0

└──▷ **BREAKING ON UPGRADE**

! Delivery confirmation is now sourced from the native execution history rather than the graph output: runs that previously resolved as `delivery_unconfirmed` because the graph output lacked the exact native receipt may now resolve differently, and any code that inspected graph output to confirm delivery must be updated to use `status.metadata.deliveryReceipt` instead.

WAS THIS USEFUL? 👍 👎

codealta

Feedback ⓘ 3.4.0 NOTES ↗ RANK

Agentic Development Environment

> CodeAlta's latest releases add mid-conversation worktree switching, Azure DevOps pull request review support, external skill folders, and a reordered phone timeline.

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

01 **Session worktree switching mid-conversation** NEW 80

Move a running session between worktrees without restarting the conversation using Change worktree... in the Workspace chip, or let agents do it with `alta session set_worktree`.

Move a running session to a feature branch worktree when you realise the work needs its own branch.

```
$ alta session set_worktree
```

— Names UI action and CLI command with a runnable example. 3.4.0

02 **Disable OpenRouter attribution headers** NEW 65

Set `request.remove_headers` in the model provider configuration to stop OpenRouter attribution headers from being sent with requests.

— Exact config key given, but no further mechanism explained. 3.4.0

03 **Azure DevOps pull request browsing** NEW 65

Browse Azure DevOps pull requests in the Issues tab, filtered by review status, vote, author, target branch, or draft state, across a repository, project, or organization.

— Lists concrete filters and scope, UI path given. 3.4.0

THINNER COVERAGE BELOW

04 **External skill folders in Settings** NEW 55

Add external skill folders in Settings › Skills › Skill folders, scoped to every space, a single space, or a single project.

— UI path and scoping options named, no deeper mechanism. 3.4.0

05 **Customizable pull request review start prompt** NEW 50

Customize what a pull request review session is asked first using the Pull request review start prompt setting in Settings › Agent prompts.

— Names setting location but lacks mechanism detail. 3.4.0

06 **Latest-first phone timeline with timestamps** IMPROVED

35

The phone timeline now shows messages in latest-first order with their timestamps, making it easier to scan recent activity at a glance.

— Describes behavior change, no named surface or command. 3.3.1

WAS THIS USEFUL?  

ogcode

Feedback ⓘ v0.58.1 NOTES ↗ RANK

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

> ogcode's latest releases focus on cutting redundant network calls with a new web-search caching and memory layer, alongside UI polish for settings and a first-party analytics proxy with an opt-out.

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

01 **First-party analytics proxy with opt-out** NEW 85

ogcode now routes analytics through a first-party proxy at `a.ogcode.in`. Users can opt out of log shipping while keeping the proxy active by setting `DO_NOT_TRACK=1` in `.env`.

Opt out of log shipping while keeping the first-party analytics proxy active.

env

```
DO_NOT_TRACK=1
```

— Named domain and exact env var/config path make it directly actionable. v0.58.0

02 **Two-tier web search response cache** NEW 70

Searches and page fetches are now served from a two-tier cache (an in-memory LRU layer over SQLite) with per-kind TTLs, so repeated reads and identical queries skip the network entirely. Cache status and a manual clear option are exposed under the search settings page.

— Mechanism and UI control described; no exact TTL values given. v0.58.1

03 **Deep-search answer memoization** NEW 65

Deep searches now memoise their synthesised answer, keyed by query, model, provider, and UTC day; a cache hit appends a served-from-cache note to the response.

— Clear key and behavior, but no manual trigger or setting named. v0.58.1

THINNER COVERAGE BELOW

04 **Per-session deep-search memory** NEW 55

Conversations now carry their own deep-search history: every turn's deep searches are recorded into the session memory file automatically at turn end, deduplicated by query.

— Behavior described but file location and format not named. v0.58.1

05 **Resizable, collapsible settings sidebar** IMPROVED 55

The settings sidebar is now resizable and folds to a 48 px icon rail.

— Exact collapsed width given; otherwise a short UI note. v0.58.0

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Code

Feedback ⓘ sdk-typescript-v0.1.27

NOTES ↗ RANK

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

> Qwen Code's two November releases add multi-modal input, session export, a concurrent batch runner, experimental LSP support and a Zed extension, plus a wave of Web Shell UI additions (sidebar activity view, split view, token usage panel, workflow chip) and new Fedora 44/RHEL 10 packaging — alongside breaking changes renaming `disable*` settings to `enable*` and removing the `read_many_files` and Smart Edit tools.

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

01 **Environment variables via settings.env** NEW 85

A new `settings.env` field lets users configure environment variables directly in settings, avoiding shell profile edits; shown with an example setting `OPENAI_API_KEY` and `HTTPS_PROXY` per-project.

Set provider API keys or proxy settings per-project without touching shell profiles.

json

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

— named config key with a runnable example sdk-typescript-v0.1.27

02 **Settings keys renamed from disable* to enable*** BREAKING 75

Settings keys previously named with a `disable*` prefix are renamed to `enable*` positive naming (e.g. `disableFoo` becomes `enableFoo`); existing config files using the old names must be updated.

— clear before/after with migration note sdk-typescript-v0.1.27

03 **prompt.submit hooks for headless runs** NEW 75

Mods can now run `prompt.submit` hooks in headless runs for `-p` prompts and `stream-json` user messages, allowing prompt inspection or rewriting before

submission.

— names hook and flags with clear mechanism v0.26.6

04 **read_many_files tool removed** BREAKING 70

The `read_many_files` tool is removed; its functionality is replaced by a `readManyFiles` utility available as a user `@-command`.

— names removed tool and its replacement sdk-typescript-v0.1.27

05 **Web Shell sidebar, split view, and workflow chip additions**

NEW 65

Web Shell adds a sidebar activity view for trusted workspace conversations (showing approval/input/error states, unread completions, running work, and bulk read/archive controls), an Add to Split View option on eligible sidebar session menus for opening sessions side by side, and a dismissible Workflow chip in the composer that turns Workflow mode off, warns when tools are unavailable, and summarizes the mode in the + menu.

— detailed UI states but no API surface UI path given, thin description behaviour described, UI only

v0.26.6

06 **Session export command** NEW 60

A new `export` command exports session history to markdown or HTML formats.

— names command and formats, no example shown sdk-typescript-v0.1.27

07 **Agent Skills stable with symlink support** IMPROVED 60

Agent Skills are promoted from experimental to stable, and the skill manager now supports symlinks so skills can be linked from external paths.

— describes change, no example or command sdk-typescript-v0.1.27

08 **Expanded ACP turns via experimental.unifiedAgentLoop** IMPROVED 60

Enables more ACP turns — including channel prompts and Goal-proposal approvals — to run on the shared loop with `experimental.unifiedAgentLoop`.

— names flag, limited mechanism detail v0.26.6

09 **Web Shell Token Usage panel enhancements** IMPROVED 60

The Web Shell Token Usage panel now shows request success rate, output speed, per-model success indicators, and distinct model colors, and plots per-request output speed, first-token latency, and failure details by model.

— metrics named, no navigation path v0.26.6

THINNER COVERAGE BELOW

10 **extra_body support for OpenAI-compatible providers** NEW 55

New `extra_body` support lets users pass provider-specific request fields to OpenAI-compatible providers.

— names field, lacks usage example sdk-typescript-v0.1.27

11 **Zed extension and experimental LSP support** NEW 55

Qwen Code adds a Zed extension for the agent server, joining the existing VS Code and JetBrains plugins, and adds experimental LSP support for code intelligence.

— names editors/plugins, no setup steps sdk-typescript-v0.1.27

12 **FORK_MODE for ProcessTransport (Electron IPC)** NEW 55

Adds `FORK_MODE` support to `ProcessTransport` for Electron IPC integration.

— names flag and integration target, no example sdk-typescript-v0.1.27

13 **contextWindowSize configuration** NEW 50

Adds `contextWindowSize` configuration support to control the model context window.

— names key, no mechanism detail given sdk-typescript-v0.1.27

14 **Concurrent runner for batch CLI execution** NEW 50

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

— no command name or flag given sdk-typescript-v0.1.27

15 **Multi-modal input across content generators** NEW 50

New multi-modal input support lets users submit images, PDFs, and audio as input across all content generators.

— describes capability, no usage steps sdk-typescript-v0.1.27

16 **Fedora 44 and RHEL 10 RPM packages** NEW 50

Desktop releases now publish Fedora 44 and RHEL 10 x64 RPM installers alongside existing Linux packages.

— names distros and package format `v0.26.6`

---

17 Smart Edit tool and ClearcutLogger removed

BREAKING

30

The Smart Edit tool and `ClearcutLogger` are removed.

— no reason or replacement given `sdk-typescript-v0.1.27`

└──▷ BREAKING ON UPGRADE

! Settings keys previously named with a `disable*` prefix are renamed to `enable*` positive naming (e.g. `disableFoo` becomes `enableFoo`); existing config files using the old names must be updated.

! The `read_many_files` tool is removed; its functionality is replaced by a `readManyFiles` utility available as a user `@ -command`.

! The Smart Edit tool and `ClearcutLogger` are removed.

WAS THIS USEFUL? 👍 👎

claude-mem

Feedback ⓘ v15.9.0 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 added Devin CLI/Desktop support, let teams point the worker at a remote host, gave My Work refresh a per-provider model setting and an on-demand button, and widened the list of credentials it redacts before storage.

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

01 **Remote worker URL configuration** NEW 88

Point the worker at a remote host with the new `CLAUDE_MEM_WORKER_URL` setting; any non-localhost URL must use https, and an invalid value now fails with a clear error instead of falling back silently.

Route all worker traffic to a shared team instance over a secure connection.

```
$ export CLAUDE_MEM_WORKER_URL=https://claude-mem-
worker.internal.example.com
```

— Exact env var, https requirement, failure behaviour, usage example v15.9.0

02 **Per-provider model for My Work refresh** NEW 85

Set the My Work refresh model per provider with the `CLAUDE_MEM_WORK_VIEW_CLAUDE_MODEL`, `CLAUDE_MEM_WORK_VIEW_CODEX_MODEL`, and `CLAUDE_MEM_WORK_VIEW_GEMINI_MODEL` environment variables, so each backend can use a different model for the automatic refresh cycle.

Use a faster or cheaper model for My Work refresh on Claude while keeping a different model for Gemini-backed projects.

```
$ export CLAUDE_MEM_WORK_VIEW_CLAUDE_MODEL=claude-3-5-haiku-
20241022
export CLAUDE_MEM_WORK_VIEW_GEMINI_MODEL=gemini-2.0-flash
```

— Names all three env vars with a runnable example v15.9.0

03 **Devin CLI and Desktop support** NEW 80

Running `npx claude-mem install` now sets up Devin CLI and Devin Desktop: hooks are written automatically, Devin's tool names are recognized, and memory search tools are registered.

— Names exact install command and what it configures v15.9.0

04 **Expanded credential redaction before storage** IMPROVED

60

More credential types are now redacted before storage: Slack and Discord webhook URLs, Stripe webhook secrets, PyPI, SendGrid, DigitalOcean, Shopify, Docker Hub, and GitLab tokens, Basic credentials after `Authorization` headers, and CLI flags like `--auth-token` and `--client-secret`.

— Lists many named credential types, no user-facing control v15.9.0

THINNER COVERAGE BELOW

05 **On-demand refresh for My Work and Past Focus** NEW 40

My Work and Past Focus now have a `Refresh now` button for on-demand refresh without waiting for the automatic cycle.

— Names UI control but gives no further mechanism v15.9.0

06 **contentSessionId on semantic injection** IMPROVED 30

Server-mode hooks now send `contentSessionId` on semantic injection.

— Bare mechanism name with no further explanation v15.9.0

WAS THIS USEFUL? 👍 👎

harness-remote

Feedback ⓘ v3.8.0 NOTES ↗ RANK

Native-session control plane for Codex, Claude Code, OpenCode, OMP and PI. Run, resume and hand off coding sessions across your machines.

> Harness Remote v3.8.0 adds a DeepSeek ACP integration for native session control across desktop, web and Android, and tightens plugin security by no longer passing gateway credentials to plugin processes.

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

01 **DeepSeek ACP integration for session control** NEW 75

Harness Remote adds a DeepSeek ACP integration, via an installed `dsh-acp` adapter or a fetched pinned package, requiring Node.js 22.15+ and DeepSeek credentials, to discover existing DeepSeek Sessions, replay conversations, and support prompts during a running turn from desktop, web, or Android.

— Names adapter, runtime requirement and capabilities; no setup command shown. v3.8.0

THINNER COVERAGE BELOW

02 **Plugin processes no longer receive gateway credentials** 30

BREAKING

Gateway credentials are no longer passed to plugin processes, isolating plugin credential scope.

— States the change but no migration detail or affected surfaces. v3.8.0

WAS THIS USEFUL? 👍 👎

pstack-claude

Feedback ⓘ v0.9.119 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's v0.9.119 release adds a mechanism for defining custom agent types via markdown files, tightens which fields those files can control, and introduces session routing for child-agent permission prompts.

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

01 **User-defined agent types from .md files** NEW 88

Placing a `.md` file in the `pstack-agents` directory inside the Pi agent folder registers a new agent type under its filename (e.g. `poteto-agent.md` adds `poteto-agent`), with no `pstack:` prefix; hidden files, unreadable entries, and names containing a colon are skipped, and a user file never overrides an existing type. An agent file's `effort:` field now acts as the default effort level when a call omits `effort`, and setting `model: inherit` runs that agent on the parent's model, though the call's own `model` and `effort` always take precedence.

– Names exact directory, file-naming rule, and config fields. v0.9.119

02 **Agent files no longer control tools, disallowedTools, or isolation**

81

BREAKING

The extension now ignores a user agent file's `tools:`, `disallowedTools:`, and `isolation:` fields; callers must instead pass `readonly: true` or `isolation: "worktree"` directly on the call.

– Exact fields and replacement call parameters given. v0.9.119

03 **PI_SUBAGENT_PARENT_SESSION routing for child permission prompts**

NEW 78

Each child `pi` process now receives `PI_SUBAGENT_PARENT_SESSION` set to the session that started it, so that with `@gotgenes/pi-permission-system` installed globally, permission prompts from child agents are routed to the parent session rather than being logged as user refusals.

– Names env var and dependent package but no setup steps shown. v0.9.119

WAS THIS USEFUL?  

Earendil Works Pi

Feedback ⓘ v1.5.0 NOTES ↗

Pi is an AI coding agent toolkit that provides a unified LLM API, agent loop, TUI interface and CLI.

Pi v1.5.0 adds custom self-update commands via `PI_SELF_UPDATE_COMMAND` or `selfUpdateCommand`, extension changelogs via `pi.changelogPath`, and new pi.exec() `maxBuffer` and `truncated` controls.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch v1.5.0 https://github.com/earendil-works/pi.git
# already have the repo? check out this version:
$ git checkout v1.5.0
```

Set `PI_SELF_UPDATE_COMMAND` or the global `selfUpdateCommand` config key so `pi update` routes through a wrapper or alternative package manager.

Add `pi.changelogPath` to an extension package's `package.json` to show a changelog after an update; use `collapseChangelog` to condense notices.

Control pi.exec() output with a `maxBuffer` option; the result now includes a `truncated` field.

MCP OAuth sign-in now defaults to `oauth.clientRegistration: "auto"`; `"dcr"` always registers dynamically; `pi mcp add --oauth-client-registration` sets the mode.

Access `Usage.cost.totalFromProvider` in SDK and extension usage data to distinguish billed totals from catalog estimates.

+1 more

NEW SUPPORTED PLATFORM GPT-6.1 Terra model support

└──▷ **BREAKING ON UPGRADE**

! TypeBox features outside JSON Schema (Type.Refine(), Type.BigInt(), Type.Undefined(), codecs, Format.Set()) no longer apply to tool arguments after migration to `@earendil-works/pi-schema`.

! The default OpenAI model changed from `gpt-5.5` to `gpt-6.1-sol`.

WAS THIS USEFUL? 👍 👎

octo-agent

Feedback ⓘ v1.17.22 NOTES ↗ RANK

Open-source, single-binary, self-hosted AI agent — your models and data stay on your machine. A coding agent on par with Claude Code and a personal assistant lighter than OpenClaw, across eight interfaces (TUI, CLI, Web, desktop, IM, VS Code, Obsidian, Mobile), all in one MIT-licensed Go binary running on any model.

> octo-agent's only change this window was a small web UI improvement: tool call cards now display their full arguments when expanded.

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

01 **Full tool-call arguments shown in web UI** IMPROVED 40

When a tool card is expanded in the web UI, octo-agent now shows the tool call's full arguments.

— Names the UI surface but no screenshot or further detail v1.17.22

WAS THIS USEFUL? 👍 👎

forgedock

Feedback ⓘ v1.43.0 NOTES ↗ RANK

ForgeDock — autonomous AI dev pipeline for Claude Code. GitHub as agent memory: issue in, PR out.

> ForgeDock's latest release adds automatic batching of child issues spawned from epics and lets teams override the default pytest guard command on a per-extension basis, both configured in `forge.yaml`.

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

01 **Automatic child-issue batching for epics** NEW 75

Setting `orchestration.auto_add_children` in `forge.yaml` automatically adds epic and decomposition child issues to a running batch as they are created, removing the need to manually re-run `forgedock orchestrate`.

Automatically enqueue child issues of an epic as they are created, without re-running `forgedock orchestrate` manually.

yaml

```yaml
orchestration:
  auto_add_children: true
```

— Names exact config key and shows runnable yaml example. v1.43.0

THINNER COVERAGE BELOW

02 **Per-extension guard-test command override** NEW 50

A per-extension guard command can now be set in `forge.yaml` to replace the default `python -m pytest` fallback used for guarded base tests.

— Describes mechanism but omits exact config key name. v1.43.0

WAS THIS USEFUL? 👍 👎

Sourcegraph Amp

Feedback ⓘ Artifacts (2026-11-02) BLOG ↗

RANK

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

> Amp introduces Artifacts, a shared document type inside threads that lets agents, users, and teammates co-read and co-edit plans, reports, and notes.

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

01 **Artifacts shared documents in threads** NEW 63

Amp adds Artifacts, a document type tied to a single thread that the agent, the user, and teammates can all read and edit, opened from Views → Artifacts in any thread. Agents now write plans, reports, and call notes as Artifacts rather than long chat replies, and read back any edits before continuing work.

Open the Artifacts panel in an active thread to read, edit, or check off items in a document the agent produced.

📍 Views → Artifacts

— Names exact UI path but lacks limits or format detail. Artifacts (2026-11-02)

WAS THIS USEFUL? 👍 👎

Cotool

Feedback ⓘ v0.78.0 NOTES ↗ RANK

Cotool is an AI security operations platform for investigating alerts, enriching threat intelligence, and automating response workflows.

> Cotool's v0.78.0 release adds a custom OpenAPI/Swagger integration builder, expands managed log ingestion to seven new sources plus SentinelOne with threat-response actions, and extends agent capabilities across alert workflows, Jira, Google Workspace, and the MCP action set.

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

01 **Custom OpenAPI/Swagger integration builder** NEW 85

Custom integrations can now be built directly from OpenAPI and Swagger specifications, with configurable authentication (API key, OAuth, or none), selectable operations to expose, and per-operation approval requirements. Configured via Integrations > Add Integration > OpenAPI / Swagger in the console, pasting the spec URL or uploading the file.

Wire up an internal ticketing or threat-intel service that has no built-in connector, using its OpenAPI spec with API-key auth.

📍 In the Cotool console, go to Integrations › Add Integration, select OpenAPI / Swagger, paste the spec URL or upload the file, configure authentication (API key, OAuth, or none), choose which operations to expose, and set approval requirements per operation.

— Names auth options, approval mechanism, and exact setup steps. v0.78.0

02 **SentinelOne ingestion with threat and alert actions** NEW 80

New `SentinelOne` managed log ingestion source adds threat and alert investigation actions for agents. Enabled via Integrations > Managed Log Ingestion > SentinelOne by entering the API URL and token and toggling Threat and alert investigation actions.

Add SentinelOne as a managed log source so agents can investigate threats and take alert actions directly from Cotool.

📍 In the Cotool console, go to Integrations › Managed Log Ingestion, select SentinelOne, enter your SentinelOne API URL and token, and enable Threat and alert investigation actions.

— Names mechanism and gives exact configuration steps. v0.78.0

03 **Managed log ingestion expanded to seven new sources** NEW

65

Managed log ingestion now supports `AWS`, `GCP`, Cloudflare, `CrowdStrike Falcon Data Replicator`, Tines audit logs, `GitHub` audit logs, and `1Password` audit events. Configured via Integrations > Managed Log Ingestion in the console; for example, 1Password ingestion requires an Events API token.

Enable 1Password audit event ingestion so agents can correlate vault access events with active investigations.

> 📍 In the Cotool console, go to Integrations › Managed Log Ingestion, select 1Password audit events, provide your 1Password Events API token, and save.

— Lists all named sources with one concrete setup example. v0.78.0

THINNER COVERAGE BELOW

04 **Alert workflow severity and Slack closure handling** IMPROVED

55

Alert workflows now let agents set severity, close true-positive alerts from Slack requests, and keep trigger alerts in one Slack thread through closure.

— Names three workflow behaviours but no config surface. v0.78.0

05 **Cotool MCP action set expanded** IMPROVED

45

Cotool MCP now includes Hunt, threat intelligence, detection, ingest health, alert routing, and agent follow-up actions.

— Names new action categories but no usage detail. v0.78.0

06 **Jira required fields on create and update** IMPROVED

45

Jira actions now support setting required fields when creating issues and updating existing issues.

— Clear behaviour but no field names or endpoint given. v0.78.0

07 **Background sub-agents sharing parent sandbox** IMPROVED

45

Sub-agents can now run in the background while sharing the parent chat's sandbox during delegated work.

— Explains mechanism but no concrete usage path given. v0.78.0

08 **ArmorCode and HaveIBeenPwned integrations** NEW

30

New integrations added: `ArmorCode` for security finding management and `HaveIBeenPwned` for breach exposure checks.

— Named integrations but no setup or mechanism given. v0.78.0

09 **Google Workspace Alert Center access for agents** IMPROVED

30

Google Workspace domain-wide delegation now includes Alert Center access for agents.

— States the capability but no further mechanism or setup. v0.78.0

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

20

GPT-6.1 Terra is now available in the model catalog.

— Bare name, no further detail on capability or use. v0.78.0

WAS THIS USEFUL? 👍 👎

Amazon Kiro

Feedback ⓘ

CLI: Automatic V2 Agent Upgrades and ACP Sign-In (2026-10-30) BLOG ↗ RANK

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

> Kiro's CLI release adds automatic V2-to-V3 agent migration with an opt-out and backups, a runtime directory override, ACP terminal sign-in, and polish to Mermaid rendering and V3 approval UI.

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

01 Automatic V2-to-V3 agent upgrade with opt-out NEW

90

Kiro now automatically migrates V2-format agents to V3 when a V3 session starts; this can be disabled by running `kiro-cli settings chat.enableAutoAgentUpgrade false`, which instead reports a count of pending upgrades. Agents in `KIRO_AGENT_CONFIG_DIR` (when set) are covered alongside workspace and global agent directories, and every rewritten file gets a `.json.bak` backup. A specific directory can also be migrated manually with `kiro-cli agent upgrade ./agents`.

Keep all agents in V2 format and get a count of pending upgrades on every V3 launch instead of migrating automatically.

```
$ kiro-cli settings chat.enableAutoAgentUpgrade false
```

Migrate a specific V2 agent directory to V3 format after a non-interactive run stops with a V2-format error.

```
$ kiro-cli agent upgrade ./agents
```

— Names exact command, env var, backup behavior, and manual migration command

CLI: Automatic V2 Agent Upgrades and AC…

02 KIRO_RUNTIME_DIR runtime unpack override NEW

65

Setting `KIRO_RUNTIME_DIR` to an absolute path controls where Kiro's bundled runtime unpacks, instead of defaulting to the home directory — useful in hardened or shared environments where home directory writes are restricted.

Unpack the bundled runtime to a controlled directory (useful in hardened or shared environments where writing to the home directory is restricted).

```
$ export KIRO_RUNTIME_DIR=/opt/kiro/runtime
kiro-cli acp
```

— Exact env var and example usage given, but no further mechanism detail

CLI: Automatic V2 Agent Upgrades and AC…

THINNER COVERAGE BELOW

03 **ACP terminal sign-in via kiro-cli login** NEW 55

ACP clients that declare terminal authentication now offer a Log in to Kiro method that runs `kiro-cli login` in a terminal when `kiro-cli acp` starts while the user is signed out.

— Names the command and trigger condition but no user-run steps

CLI: Automatic V2 Agent Upgrades and AC…

04 **V3 approval prompt and rule page UI tweaks** IMPROVED 55

V3 approval prompts now show only the first line of a multi-line command, with Ctrl+O to expand the rest; the V3 approval rule page now highlights scope choices with an `s` hint and names the tool (such as Shell) in the entire-tool row.

— Gives a concrete keyboard shortcut and UI labels but limited mechanism

CLI: Automatic V2 Agent Upgrades and AC…

05 **Mermaid flowchart rendering improvements** IMPROVED 45

Mermaid flowcharts now render node shapes with type colors, draw nested subgraphs with group markers, and fit long labels within narrow terminals.

— Describes visual changes but nothing for the reader to invoke

CLI: Automatic V2 Agent Upgrades and AC…

WAS THIS USEFUL? 👍 👎

AGENT

◆ MCP Servers & Tooling

hayhooks

Feedback ⓘ v2.2.0 NOTES ↗ RANK

Easily deploy Haystack pipelines as REST APIs and MCP Tools.

> Hayhooks adds a way to lock down production deployments by disabling runtime pipeline deploy/undeploy across HTTP, CLI, and MCP, while still serving pipelines loaded at startup.

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

01 **Runtime deploy lockdown via environment flag** NEW 93

Setting `HAYHOOKS_RUNTIME_DEPLOY_ENABLED=false` disables runtime deploy and undeploy across HTTP, CLI, and MCP, while pipelines loaded from `HAYHOOKS_PIPELINES_DIR` at startup remain executable. When disabled, HTTP deployment routes return `404` and MCP omits the `deploy_pipeline` and `undeploy_pipeline` tools, rejecting direct calls to them.

Lock down a production server so operators cannot push new pipeline code at runtime, while still serving pipelines baked into the image or volume at startup.

```
$ export HAYHOOKS_PIPELINES_DIR=/app/pipelines
export HAYHOOKS_RUNTIME_DEPLOY_ENABLED=false
hayhooks run
```

— Names exact env var, HTTP status, MCP tool names, and runnable example v2.2.0

WAS THIS USEFUL? 👍 👎

api-mcp-server

Feedback ⓘ v3.25.0 NOTES ↗ RANK

MCP server that exposes Hostinger's platform API for hosting deployment, CDN control, domain management, and WordPress operations.

> api-mcp-server added domain-alias management operations and a new website creation operation for Hostinger's hosting API in this release.

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

01 **Parked domain and subdomain management operations** NEW

70

Three new operations let users manage parked (alias) domains and subdomains on a hosted website: `hosting_websites_list-domains` to list them, `hosting_websites_add-domain` to add a parked domain or subdomain, and `hosting_websites_remove-domain` to remove one.

— Names three exact operations but no further mechanism detail v3.25.0

THINNER COVERAGE BELOW

02 **Website creation operation (deprecated variant)** DEPRECATED

40

Adds a `hosting_websites_create-deprecated` operation for creating a website.

— Only the operation name is given, no context on deprecation reason v3.25.0

WAS THIS USEFUL? 👍 👎

kordoc

Feedback ⓘ v4.24.41 NOTES ↗ RANK

모두 파싱해버리겠다 — HWP·HWPX·PDF·Office 문서를 Markdown으로. 양식 자동 채우기와 신구대조를 갖춘 CLI·MCP 서버 | Convert Korean documents (HWP, HWPX, PDF, Office) to Markdown — CLI and MCP server with form filling and diff

> kordoc's latest releases focus on deeper HWP-to-HWPX conversion fidelity: previously-rejected protected files can now be converted, change-tracking annotations are preserved, a native Hancom save engine option was added for Windows, and formula regions can be OCR'd into LaTeX.

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

01 Hancom native save engine for HWPX conversion NEW

90

Documents can be converted to HWPX using the Hancom application's native save engine on Windows via `--engine hancom|auto` (CLI), the `{ engine }` option in convertToHwpx(..., { engine }) (API), or the `engine` parameter on the MCP `convert_to_hwpx` tool, preserving HWP 3.x and HWPML formatting without triggering confirmation, password, or version-warning dialogs.

Use the Hancom native save engine when converting a legacy HWP file to HWPX on a Windows machine that has Hancom installed, to preserve formatting that the built-in engine may not fully replicate.

```
$ npx kordoc to-hwpx 문서.hwp --engine hancom -o 문서.hwpx
```

— Named flags/options across CLI, API, MCP plus a runnable command v4.24.38

02 Change-tracking data in HWP-to-HWPX conversion NEW

80

HWP-to-HWPX conversion now carries change-tracking data from `ViewText`: delete regions tagged with `deleteBegin` and `deleteEnd` (range number and change number), character shape changes via run `charTcId`, paragraph shape changes via `paraTcId`, and a header-level change list in `trackChanges` (type, timestamp, author, hidden flag, shape number) plus `trackChangeAuthors`.

— Mechanism and field names given but no direct invocation example v4.24.41

03 Convert protected HWP files to HWPX NEW

75

Distribution-locked HWP files, which were previously rejected, can now be converted to HWPX via `hwpToHwpx`, `convertToHwpx`, the CLI `to-hwpx` subcommand, or the MCP `convert_to_hwpx` tool.

Convert a distribution-locked HWP file (previously rejected) to HWPX while preserving change-tracking annotations.

```
$ npx kordoc to-hwpx 배포용문서.hwp -o 배포용문서.hwpx
```

— Names all four invocation surfaces and the before/after behavior v4.24.41

THINNER COVERAGE BELOW

04 **Formula OCR to LaTeX** NEW 55

Running `formulaOcr` replaces math-region text with `$LaTeX$` equivalents when processing image-input documents.

— Names the function and output format but no usage example v4.24.38

UPDATED DETECTION OCR accuracy improvements for rotated pages, two-column layouts, statistical tables, and official gazettes

v4.24.39 NOTES ↗ STABLE

kordoc now automatically rotates landscape-scanned pages before OCR grid-line detection; normal pages are unaffected.

└──▷ GET THIS VERSION ▾

npm

```
$ npm install -g kordoc@4.24.39
```

git

```
$ git clone --branch v4.24.39
https://github.com/chrisryugj/kordoc.git
# already have the repo? check out this version:
$ git checkout v4.24.39
```

named in the project's README · checked 3 Nov Wrong command?

WAS THIS USEFUL? 👍 👎

premiere-pro-mcp

Feedback ⓘ v1.29.0 NOTES ↗ RANK

Adobe Premiere Pro MCP — independent, local-first server for supported workflows. Connect Claude, Codex, or Cursor through a local CEP bridge.

> premiere-pro-mcp v1.29.0 adds startup controls for trimming and compacting the tool schema, safer chunked bulk deletion, versioned timeline snapshots, and guarded bin-restructuring plans for project audits.

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

01 Chunked bulk deletion in delete_multiple_project_items

IMPROVED 90

`delete_multiple_project_items` now accepts up to 500 `item_ids`, deletes in chunks of 25 items per host command with per-chunk readback, sends progress notifications at least every 10 seconds, and reports partial outcomes via `DELETE_CHUNK_STOPPED` instead of claiming success.

— Specific limits, chunk size, notification cadence, and failure signal. v1.29.0

02 Versioned snapshots in inspect_timeline_snapshot IMPROVED

83

`inspect_timeline_snapshot` can now write a versioned JSON snapshot inside an approved private workspace without overwriting an existing file. Setting `return_content: false` returns bounded identity, counts, completeness, byte length, and SHA-256 instead of clip content.

— Names the tool, the field, and the exact returned metadata. v1.29.0

03 Explicit tool selection via PREMIERE_MCP_TOOL_LIST NEW

80

Select an explicit subset of tools at startup using the `PREMIERE_MCP_TOOL_LIST` environment variable, which unions with selected packs, preserves discovery tools, and rejects unknown names. Reviewable lists in `fork/packs/*.txt` support focused client configurations.

Limit an MCP stdio process to a focused set of tools to reduce schema overhead in an automated pipeline.

```
$ PREMIERE_MCP_TOOL_LIST=inspect_timeline_snapshot,delete_multiple_project_items PREMIERE_MCP_SCHEMA_PROFILE=compact node server.js
```

— Names exact env var, pack file path, and union/reject behavior. v1.29.0

04 **Compact schema mode via PREMIERE_MCP_SCHEMA_PROFILE** NEW

75

Setting `PREMIERE_MCP_SCHEMA_PROFILE=compact` enables compact schema mode, which adds the memory-only `set_session_guard` tool for a single stdio process; stateless HTTP refuses compact mode.

Limit an MCP stdio process to a focused set of tools to reduce schema overhead in an automated pipeline.

```
$ PREMIERE_MCP_TOOL_LIST=inspect_timeline_snapshot,delete_multiple_project_items PREMIERE_MCP_SCHEMA_PROFILE=compact node server.js
```

— Exact config key and value given, plus the restriction on HTTP mode. v1.29.0

THINNER COVERAGE BELOW

05 **Guarded bin planning in audit_project_structure** IMPROVED

45

`audit_project_structure` now accepts a target tree with guarded bin placement, rename, and empty-bin removal plans.

— Names the tool but gives no mechanism for how guarding works. v1.29.0

WAS THIS USEFUL? 👍 👎

context7 Feedback ⓘ @upstash/context7-mcp@4.5.0 NOTES ↗ RANK

Context7 maintains up-to-date documentation and context for AI agents through an MCP server.

> Context7 added a research-preview feedback tool for rating documentation answers and shipped a new `ctx7 search` command for cross-library documentation lookup, alongside fixes to how `ctx7 setup` / `remove` handle OpenCode's config layout.

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

01 Docs feedback tool in research preview NEW 87

Enabling the research preview with `--research-preview` or `CONTEXT7_RESEARCH_PREVIEW=1` (stdio) or `?research_preview=1` on the MCP URL (hosted server) unlocks a `submit-docs-feedback` tool that lets agents rate `query-docs` results with a score from 0 to 5, a description of what was wrong, and a short comment. Research-preview API calls now carry the `X-Context7-Research-Preview: 1` header, and the server instructs the agent to call `submit-docs-feedback` automatically after answering.

Run the context7 MCP server in stdio mode with research preview on so the `submit-docs-feedback` tool is listed and agents are prompted to rate every answer.

```
$ CONTEXT7_RESEARCH_PREVIEW=1 npx @upstash/context7-mcp
```

Connect to the hosted context7 MCP server with research preview enabled via the query string.

```
$ https://mcp.context7.com/mcp?research_preview=1
```

— Names exact flags, env var, header and scoring mechanism @upstash/context7-mcp@4.5.0

02 Cross-library documentation search command NEW 80

`ctx7 search <query>` queries documentation across all libraries in one step without needing a library ID, with optional `--library`, `--library-version`, and `--language` flags to narrow results, and `--json` for structured output.

Find documentation for a specific topic across all libraries without knowing the library ID first — useful when you are exploring an unfamiliar ecosystem.

```
$ ctx7 search "rate limiting" --language typescript --json
```

Narrow a cross-library search to a specific library and version when you know the library but still want a one-step query.

```
$ ctx7 search "authentication" --library redis --library-version
5
```

— Runnable command with all flags named `ctx7@0.9.0`

THINNER COVERAGE BELOW

03 **OpenCode config layout handling in ctx7 setup/remove** IMPROVED

54

`ctx7 setup` now writes the OpenCode MCP entry inside `mcp.servers` when the config uses the OpenCode 2.x layout, and moves entries an earlier version wrote next to `mcp.servers`; `ctx7 remove` finds Context7 in either the 1.x or 2.x layout.

— Names config key and layouts but no command example `ctx7@0.9.0`

WAS THIS USEFUL? 👍 👎

mcp-server

Feedback ⓘ v2.5.0 NOTES ↗ RANK

BrowserStack's Official MCP Server

> BrowserStack's MCP Server added a new `ask-browserstack` feature that connects BrowserStack AI agents to the server.

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

01 **BrowserStack AI agents via ask-browserstack** NEW

40

Adds support for BrowserStack AI agents through the `ask-browserstack` feature.

— Names the feature but not its mechanism or usage. `v2.5.0`

WAS THIS USEFUL? 👍 👎

godot-ai

Feedback ⓘ v4.4.0 NOTES ↗ RANK

godot-ai is an MCP server and AI tools for the Godot game engine.

> godot-ai added a scene tab close operation to its scene management tool.

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

01 **Close operation in scene_manage** NEW 65

The `scene_manage` tool gains a new `op="close"` option to dismiss an open scene tab programmatically.

— Names exact tool and flag but little mechanism detail v4.4.0

WAS THIS USEFUL? 👍 👎

novamira

Feedback ⓘ v1.12.14 NOTES ↗ RANK

MCP server that gives AI agents full access to WordPress through PHP execution and filesystem operations

> novamira's v1.12.14 release expands how AI tools connect to its Configuration page, adding broader OAuth/Application Password support, six new CLI-ready AI tools, and a clearer HTTPS requirement notice.

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

01 OAuth and Application Password auth for AI tools NEW

67

AI tools on the Configuration page that previously required the Novamira CLI — such as ZCode — now also support OAuth and Application Password connection methods. For AI tools not on the built-in list, a new Other MCP client option provides the server URL for OAuth or a prompt to paste for Application Password.

Connect an AI tool not in novamira's built-in list by grabbing the OAuth server URL or Application Password prompt from the Other MCP client option.

📍 In the Configuration page, scroll to the AI tools section, select Other MCP client, then copy the server URL (for OAuth) or the paste prompt (for Application Password) into your client.

— Names auth methods and UI path but lacks deeper mechanism v1.12.14

THINNER COVERAGE BELOW

02 Novamira CLI setup for six new AI tools NEW

55

The Configuration page now supports setting up the Novamira CLI for six new AI tools: Grok Build, MiniMax Code, Posit Assistant, Kimchi, fx, and Sarvam Code.

— Lists all named tools but no mechanism beyond listing v1.12.14

03 HTTPS requirement notice for CLI on non-HTTPS sites IMPROVED

30

On sites without HTTPS, the Configuration page now shows the Novamira CLI as requiring HTTPS instead of omitting it entirely.

— Simple UI behavior change, minimal detail v1.12.14

WAS THIS USEFUL? 👍 👎

pullmd

Feedback ⓘ v3.17.0 NOTES ↗ RANK

Self-hosted URL- and file-to-Markdown service for humans and AI agents - web pages, documents, images, audio, YouTube. PWA + REST + MCP + Claude Code skill, Reddit-aware, refreshable share links.

> pullmd v3.17.0 adds a health-check capability that surfaces outdated Playwright sidecars to plain-HTTP monitors.

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

01 Degraded status reporting for outdated Playwright sidecars

83

BREAKING

The `GET /api/status` endpoint now reports a `pullmd-playwright` sidecar that is missing the `ssrf` policy (or predates v3.15.2) as `{"status": "degraded", "error": "policy-missing"}` with a `503` response, making the version mismatch visible to plain-HTTP monitors such as Uptime Kuma. Renders are not blocked when this occurs. Note: an instance that updates `pullmd` but keeps a `pullmd-playwright` image older than v3.15.2 will now receive `503` from `GET /api/status` until the sidecar is also updated.

— Names endpoint, response body and status codes; includes upgrade note v3.17.0

└──▷ **BREAKING ON UPGRADE**

! An instance that updates `pullmd` but keeps a `pullmd-playwright` image older than v3.15.2 now receives `503` from `GET /api/status` until the sidecar is also updated.

WAS THIS USEFUL? 👍 👎

mcp-grafana

Feedback ⓘ v2.2.0 NOTES ↗ RANK

MCP server for Grafana

> mcp-grafana v2.2.0 adds CloudWatch Logs Insights tools, an opt-in Prometheus query guardrail, BigQuery support for its SQL tools, and exposes datasources as MCP resources, alongside a migration of datasource queries to `/api/ds/query` that breaks several existing tool outputs.

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

01 **Prometheus query guardrail** NEW 95

An opt-in guardrail applies to `query_prometheus`, `query_prometheus_histogram`, and Prometheus panels in `run_panel_query`, configured via `--prometheus-guardrail-mode=shadow|enforce`, `--prometheus-guardrail-max-range` (default `192h`), and `--prometheus-guardrail-max-points` (default `11000`). In enforce mode, over-budget queries are rejected with rewrite guidance.

— Exact flags, defaults and enforcement behaviour given v2.2.0

02 **Loki cost guardrail extended** IMPROVED 90

The Loki cost guardrail now also covers `query_loki_patterns`, adds `--loki-guardrail-on-estimate-failure=allow|block` to control behaviour when query cost cannot be assessed, and records an `mcp.loki_guardrail.estimated_bytes` histogram metric. The guardrail is also available as the standalone `github.com/grafana/mcp-grafana/lokiguard` Go module.

Block Loki queries whose cost cannot be estimated, so agents never fire uncosted log scans.

```
$ mcp-grafana --loki-guardrail-on-estimate-failure=block
```

— Named tool, flag, metric and standalone module with example v2.2.0

03 **CloudWatch Logs Insights support** NEW 87

Adds CloudWatch Logs Insights via three tools in the opt-in `cloudwatch` category: `list_cloudwatch_log_groups`, `list_cloudwatch_log_group_fields`, and `query_cloudwatch_logs`. Queries run through `/api/ds/query` with a server-side row limit (default 100, max 1000) and support cross-account log groups by ARN.

— Names exact tools, endpoint and limits; enabling category flag not given v2.2.0

04 **Datasource queries migrated to /api/ds/query** BREAKING

75

Datasource queries now go through `/api/ds/query` instead of the datasource proxy. Grafana before 9.0 is no longer supported for Prometheus, Loki, Quickwit, Graphite, Pyroscope, VictoriaLogs, Elasticsearch, and Tempo queries, and a token without datasource-query permission now gets Grafana's error instead of being routed around the query service.

— Clear before/after and version cutoff, but no migration command v2.2.0

---

05 **SSE keepalive for proxied connections** NEW 70

The new `--sse-keepalive` flag pings idle SSE clients at a given interval to keep connections alive through reverse proxies or load balancers; off by default.

Keep an SSE session alive through an nginx reverse proxy with a 60-second idle timeout.

```
$ mcp-grafana --sse-keepalive=30s
```

— Exact flag with runnable example, but mechanism is brief v2.2.0

---

06 **Datasources as MCP resources** NEW 70

Datasources are now exposed as MCP resources at `grafana://datasources/{uid}`, listed per caller with their own credentials, so clients with a resource picker (such as `@` in Claude Code) can attach a datasource to a conversation directly.

— Exact resource URI given; usage is UI-driven, not a command v2.2.0

---

07 **BigQuery support in SQL tools** NEW 70

Adds BigQuery support to `query_sql`, `list_sql_databases`, `list_sql_tables`, and `describe_sql_table`, implemented without querying `INFORMATION_SCHEMA` so metadata lookups do not incur billable usage.

— Names tools and the billing-avoidance mechanism v2.2.0

---

08 **Interactive Grafana App rendering via render_grafana_app** 60

BREAKING

A new `render_grafana_app` tool displays the result of an enabled query tool (`query_prometheus` or `get_tempo_trace`) in the interactive Grafana App for MCP hosts that support MCP Apps. This replaces the prior behaviour where `query_prometheus` and `get_tempo_trace` opened an App themselves, which they no longer do.

— Named tool and behaviour change, but no flag or endpoint v2.2.0

---

09 **Folder path filtering in search_dashboards** IMPROVED

60

`search_dashboards` results now include a `folderPath` field (for example `Ops/Production/Services`) and accept an optional `folderPath` input to restrict results to a folder and all its subfolders.

— Named field and input with a concrete example value `v2.2.0`

THINNER COVERAGE BELOW

10 **Tempo trace output shape change** BREAKING 55

`search_tempo_traces` (on Grafana 12.3 and later) and `get_tempo_trace` return a new output shape with millisecond times; `search_tempo_traces` no longer returns job counts, `serviceStats`, or the partial-results note.

— Specific before/after fields named, no action step given `v2.2.0`

11 **Pyroscope profile metrics now rates** BREAKING 55

`query_pyroscope` metrics for cumulative profile types are now rates (for example cores instead of CPU nanoseconds per step), with a new `unit` field, and the step is at least the datasource's `minStep`.

— Concrete before/after units and new field, no action step `v2.2.0`

12 **Infinity and Neo4j panel support in run_panel_query** NEW 50

`run_panel_query` can now run Infinity panels that use a backend parser and Neo4j panels on the `kniepdennis-neo4j-datasource` plugin.

— Names tool and plugin but describes no mechanism `v2.2.0`

13 **Graphite time format restrictions** BREAKING 50

`query_graphite` rejects Graphite absolute time formats other than Unix seconds, `YYYYMMDD`, and RFC3339; `tags.name` is now the series target.

— Names accepted formats and field change, no migration step `v2.2.0`

14 **Time-bounded Tempo trace lookup** IMPROVED 45

`get_tempo_trace` now accepts optional `start` and `end` time bounds so Tempo does not have to scan every block for the trace.

— Names the tool and fields but is a single short line `v2.2.0`

15 **Native histogram sum removed** BREAKING 25

Native histograms returned from `query_prometheus_histogram` no longer include their sum.

— Single-line change with no further context `v2.2.0`

└──▷ **BREAKING ON UPGRADE**

! Datasource queries now go through `/api/ds/query` instead of the datasource proxy; Grafana before 9.0 is no longer supported for Prometheus, Loki, Quickwit, Graphite, Pyroscope, VictoriaLogs, Elasticsearch, and Tempo queries, and a token without datasource-query permission gets Grafana's error instead of being routed around the query service.

! `search_tempo_traces` (on Grafana 12.3 and later) and `get_tempo_trace` return a new output shape with millisecond times; `search_tempo_traces` no longer returns job counts, `serviceStats`, or the partial-results note.

! `query_pyroscope` metrics for cumulative profile types are now rates (for example cores instead of CPU nanoseconds per step), with a new `unit` field, and the step is at least the datasource's `minStep`.

! `query_graphite` rejects Graphite absolute time formats other than Unix seconds, `YYYYMMDD`, and RFC3339; `tags.name` is now the series target.

! Native histograms from `query_prometheus_histogram` no longer include their sum.

! `query_prometheus` and `get_tempo_trace` no longer open an App themselves; use the new `render_grafana_app` tool to display results in the Grafana App.

WAS THIS USEFUL? 👍 👎

brightdata-mcp

Feedback ⓘ v4.3.0 NOTES ↗ RANK

A powerful Model Context Protocol (MCP) server that provides an all-in-one solution for public web access.

> brightdata-mcp shipped new parallel search and batch page-fetch tools with configurable formats and timeouts, then migrated search to a dedicated SERP API zone with recency filtering in v4.3.0 — a release that also renamed several parameters and error codes and dropped Yandex support.

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

01 Search tools with SERP zone and time-range filter NEW

96

`search_engine` runs a single search and `search_engine_batch` runs up to 5 in parallel with the same parameters, returning a JSON array with one result or error per query; `search_engine_batch` caps execution at 45 seconds, returning completed searches and a `timeout` error for the rest. Both now query a dedicated SERP API zone instead of Web Unlocker, controlled by the `SERP_ZONE` environment variable (default `mcp_serp`, auto-created on startup if missing), and return structured JSON in the form `{organic: [{link, title, description}]}` with Bing as the default engine. Results can be filtered by recency via the new `time_range` parameter: `hour`, `day`, `week`, `month`, `year`, or `any` (default); Bing supports `day`, `week`, and `year`.

Point search tools at a non-default SERP zone, such as a pre-existing zone named `my_serp_zone`, rather than letting the server create `mcp_serp`.

```
$ SERP_ZONE=my_serp_zone npx brightdata-mcp
```

— Names env var, params and output shape with a runnable example v4.3.0

02 Fetch API page-fetch tools NEW

90

The Fetch API tools `scrape_as_markdown` (single page) and `scrape_batch` (up to 5 pages) fetch pages and return JSON with `url`, `title`, `format`, and `content`; content defaults to Markdown but accepts `text`, `html`, or `ld_json` via the `format` parameter. Single-page fetches time out at 45s, batch fetches at 35s, and batch results come back as a JSON array in input order with just `url` on failure. Both tools now return standardised error codes: `rate_limited`, `service_unavailable`, `invalid_input`, `quota_exceeded`, `auth_failed`, and `site_rate_limited`.

— Names both tools, all fields, timeouts and error codes v4.3.0

03 Search parameter and zone migration

71

BREAKING

The `geo_location` parameter is renamed to `country`, and `cursor` is renamed to `page` (a number from 0, 10 results per page); the old names now return an `invalid_input` error. `search_engine` and `search_engine_batch` now require a SERP API zone (defaulting to `mcp_serp`) instead of Web Unlocker, so deployments relying on a Web Unlocker zone for search must migrate. Yandex is removed as a supported search engine.

— Names old and new param names and migration requirement v4.3.0

THINNER COVERAGE BELOW

04 **Fetch API error code renames** 38

BREAKING

Fetch API error code `busy` is renamed to `rate_limited`, `service_error` to `service_unavailable`, and `invalid_url` to `invalid_input`.

— Names old and new codes but no further mechanism v4.3.0

└──▷ **BREAKING ON UPGRADE**

! The `geo_location` parameter is renamed to `country`; the old name now returns an `invalid_input` error.

! The `cursor` parameter is renamed to `page` (a number from 0, 10 results per page); the old name now returns an `invalid_input` error.

! Fetch API error code `busy` is renamed to `rate_limited`, `service_error` to `service_unavailable`, and `invalid_url` to `invalid_input`.

! Yandex is removed as a supported search engine.

! `search_engine` and `search_engine_batch` now require a SERP API zone (defaulting to `mcp_serp`) instead of Web Unlocker; deployments relying on a Web Unlocker zone for search must migrate.

WAS THIS USEFUL?

claudexor

Feedback ⓘ v5.0.2 NOTES ↗ RANK

Multi-harness control plane for Claude Code, Codex, Cursor, and OpenCode: quota-aware rotation across multiple Claude/Codex subscriptions, shared thread context, and cross-model review.

> Claudexor's v5.0.2 release adds a new workspace.prepared run event, unifies capability-cache reads behind a shared daemon projection, streams run events directly from the event log with explicit error codes, and hardens storage of large bodies and run receipts — alongside two breaking API changes.

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

01 Shared capability cache for agent-capabilities endpoint and MCP tool

IMPROVED 90

Capability reads on `GET /v2/agent-capabilities` and the MCP `claudexor_capabilities` tool now share one daemon status-projection cache: concurrent in-window reads share one harness sweep, a computed catalog is served for 15 seconds with its build time in `generatedAt`, and credential mutations or setup logins invalidate the cache immediately.

— Names endpoint, MCP tool, TTL, and invalidation triggers precisely. v5.0.2

02 Direct event-log reads for run event streams IMPROVED

90

Run event streams on `GET /v2/runs/:id/events` now re-read the run's `events.jsonl` directly instead of queuing on the shared bulk worker, and streams end with `event: error` carrying either `run_event_not_found` (retryable) or `run_expired_by_retention` (not retryable) once the job is terminal.

— Gives endpoint, mechanism change, and exact error codes. v5.0.2

03 workspace.prepared run event in timeline and CLI BREAKING

70

A new `workspace.prepared` run event now appears on the run timeline, in the macOS turn transcript, and in CLI follow output before the executor starts, reporting the source folder, selected paths, file/directory/symlink counts, and any selected paths not found. The run event type list gains `workspace.prepared` with a generated `WorkspacePreparedPayload` schema; consumers that pin the generated JSON schemas must update them.

— Names event, payload schema, and surfaces; lacks exact schema diff. v5.0.2

04 **Folder custody on failed capture in directory mode** IMPROVED

70

An isolated folder copy in directory mode without `--in-place` is now kept under per-attempt custody when the final capture refuses (for example, on an unreadable file or a FIFO), instead of being deleted; the `discard` decision releases it.

— Names the flag and failure modes, gives clear before/after behaviour. v5.0.2

---

05 **Write-through staging for large stored bodies and run receipts**

IMPROVED

65

Large stored bodies over 64 KiB and a run's terminal receipts (`final/run_facts.yaml`, `final/telemetry.yaml`) are now staged as write-through temp files on Node's file worker pool before the database commit, so the synchronous commit only registers or renames already-existing files.

— Internal mechanism named but no user-facing action available. v5.0.2

---

06 **EventLog.flushDeferredTerminal() replaced with commitDeferredTerminal()**

65

BREAKING

EventLog.flushDeferredTerminal() is removed; callers must replace it with await log.commitDeferredTerminal().

— Exact migration path given but no further explanation of why. v5.0.2

└──▷ **BREAKING ON UPGRADE**

! EventLog.flushDeferredTerminal() is removed; callers must replace it with await log.commitDeferredTerminal().

! The run event type list gains `workspace.prepared` with a generated `WorkspacePreparedPayload` schema; consumers that pin the generated JSON schemas must update them.

WAS THIS USEFUL? 👍 👎

Composio

Feedback ⓘ announced · Oct 21, 2026 BLOG ↗

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

Composio adds Firewall, a runtime boundary for agentic software that checks agent actions and incoming tool results in both directions before execution.

NEW DETECTION Composio Firewall now evaluates every agent action through Composio against user intent and developer-set boundaries before execution, using Jev and custom small language models tuned on real-world usage data.

NEW DETECTION Composio Firewall now checks tool results before they reach the agent and blocks results carrying malicious instructions, protecting against prompt injection at the point where outside information enters the agent's context.

WAS THIS USEFUL? 👍 👎

MCP Python SDK

Feedback ⓘ CVE-2026-116684 CVE ↗ STABLE

The MCP Python SDK provides clients and servers for exposing tools, resources, and prompts through Model Context Protocol.

`CVE-2026-116684`: Server bearer-token authentication accepts tokens issued for a different resource server (CVSS 6.8, affects `>= 2.0.0, < 2.2.0`)

▾ 2 more releases in this issue · 2026-09-28 → 2026-10-05

> CVE-2026-116685 CVE ↗ STABLE
>
> `CVE-2026-116685`: HTTP client transports followed cross-origin redirects carrying custom headers and bodies (CVSS 5.9, affects `>= 2.0.0, < 2.2.0`)

> CVE-2026-116686 CVE ↗ STABLE
>
> `CVE-2026-116686`: Client fetched server-chosen $ref URLs while validating tool results against output schemas (CVSS 6.5, affects `>= 2.0.0, < 2.2.0`)

WAS THIS USEFUL? 👍 👎

◆ AI Agent Frameworks

agentlas-os

Feedback ⓘ os-v1.2.78 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 introduced a standalone `os-v` release channel with its own isolated runtime and opt-in feature toggles, while extending host integrations (Claude Code, Codex, Grok, Antigravity) with safer session isolation, automatic skill-alias migration, and headless permission checks; it also made hub agents skip unneeded project file reads and removed graph automation and Telegram/Connect commands.

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

01 Standalone release channel and engine runtime NEW

95

Agentlas OS now installs into its own runtime at `~/.agentlas/os-runtime` with the `~/.agentlas/os-bin/agentlas-os` launcher, leaving an existing Desktop runtime untouched. Optional features — project setup, project index, Context Map, memory, Storm, automatic updates, network index, and telemetry — are off by default; check status with `agentlas-os features status` and toggle with `features on|off <name>`, `storm on|off`, `memory on|off`, `context map on|off`, and `one on|off`.

Check which optional features are currently on or off before enabling memory and Storm for a session.

```
$ agentlas-os features status
agentlas-os memory on
agentlas-os storm on
```

— Exact runtime paths and toggle commands given os-v1.2.73

02 Claude Code and Codex session hosting IMPROVED

88

Connect Claude Code and Codex to the engine with `agentlas-os host session claude` or `host session codex`, and inspect what is connected with `host status`. After updating, `host session codex` now automatically moves its own Codex skill alias to the current runtime without manual intervention and never replaces an alias it does not own; separately, session hosting in Claude Code or Codex now isolates the engine's location variables so other Agentlas plugins can't downgrade the standalone runtime.

Route Claude Code through the Agentlas OS engine and confirm it is connected.

```
$ agentlas-os host session claude
agentlas-os host status
```

— Names commands and the isolation/alias mechanism but behavior is automatic os-v1.2.73 os-v1.2.74 os-v1.2.75

03 **Grok and Antigravity plugin registration** NEW 81

Register, inspect, or remove the Grok plugin via `agentlas-os host install|status|remove grok`, exposing tools as `agentlas-os__<tool>` inside Grok's plugin manager. Antigravity and Grok now share one registration module, so `host install|status|remove antigravity` continues to work as before, registering the MCP server `agentlas-os` with Antigravity's plugin manager.

Install the Agentlas OS plugin into Grok so its tools appear under the `agentlas-os__<tool>` namespace in Grok's plugin manager.

```
$ agentlas-os host install grok
```

Check whether the Grok plugin is correctly registered before running a workflow, or remove it when rotating to a different host.

```
$ agentlas-os host status grok
agentlas-os host remove grok
```

Register the agentlas-os standalone plugin with Antigravity's plugin manager so the MCP server `agentlas-os` is available inside Antigravity.

```
$ agentlas-os host install antigravity
```

Check whether the standalone plugin is currently registered in Antigravity without making any changes.

```
$ agentlas-os host status antigravity
```

— Runnable install/status/remove commands for two hosts os-v1.2.76 os-v1.2.74

04 **Headless Antigravity MCP permission check** NEW 75

Check what MCP permission rule a headless `agy -p` run requires with `agentlas-os host status antigravity`, letting operators set the rule ahead of time before running Antigravity in CI.

Before running Antigravity headlessly in CI, check which MCP permission rule is required so you can set it ahead of time.

```
$ agentlas-os host status antigravity
```

— Specific command tied to a concrete CI scenario `os-v1.2.75`

THINNER COVERAGE BELOW

05 **Context-aware project file reads for hub agents** IMPROVED

35

Hub agents now read project files only when a request concerns the open project, and leave the project alone for self-contained requests or when no project is open, reducing unnecessary file access.

— Describes behavior change but no command or flag named `os-v1.2.78`

06 **Graph automation and Telegram commands removed** BREAKING

30

Graph automation and the Telegram/Connect commands have been removed from the engine.

— Names removed features but no migration guidance given `os-v1.2.73`

└──▷ **BREAKING ON UPGRADE**

! Graph automation and the Telegram/Connect commands have been removed from the engine.

WAS THIS USEFUL? 👍 👎

lexmount Moli

Feedback ⓘ v1.1.21 NOTES ↗ RANK

Best headless browser for AI agents. Lite, Fast, High-Compatibility. Built in Rust

> Moli v1.1.21 adds durable session storage, worker geometry APIs, and native text module imports for ES modules.

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

01 **Persistent cookie and localStorage storage** NEW 55

Moli now persists cookies and localStorage in partition SQLite databases via feat(storage), enabling durable cross-session state for headless browser sessions.

— Names mechanism and feature flag but no API or config key v1.1.21

02 **Text module imports for ES modules** NEW 50

Moli supports native text module imports via feat(modules), allowing plain-text resources to be imported as ES modules.

— Names feature flag and capability but lacks usage example v1.1.21

03 **Native geometry interfaces in workers** NEW 13

Moli adds native geometry interfaces in workers, per the release summary, though no further detail on the specific APIs or usage is given.

— Only named in summary, no mechanism or detail provided v1.1.21

WAS THIS USEFUL? 👍 👎

HeyGen HyperFrames

Feedback ⓘ v0.9.7 NOTES ↗ RANK

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

> HyperFrames shipped a breaking change to cloud rendering's motion-blur flag, a stricter GSAP tween lint rule, smarter video preview fallbacks, and new Studio APIs for key-binding queries and pre-building lighter clip copies.

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

01 `--page-motion-blur` removed from cloud render

BREAKING

90

`hyperframes cloud render` no longer accepts `--page-motion-blur`; use `data-motion-blur="on"` in the composition instead. Local `hyperframes render --page-motion-blur` is unchanged.

— Exact flag, replacement, and scope of change named. v0.9.7

02 Overlapping GSAP tween lint rule extended

IMPROVED

70

The `overlapping_gsap_tweens` lint rule now also flags a `set` call that fires while a tween on the same element and property is still running — a pattern that plays differently from how it scrubs and exports.

— Names exact lint rule and new trigger condition. v0.9.7

03 Preview fallback for unplayable video formats

IMPROVED

60

Preview now shows a video's own frame at the playhead instead of black when the browser cannot play the format (such as HEVC without a decoder or ProRes) while its playable copy is being made, and copy progress is now reported at Full quality too.

— Describes behavior and named formats but no UI path given. v0.9.7

THINNER COVERAGE BELOW

04 Studio API for key combination queries

NEW

50

Studio hosts can now query which action holds a key combination, letting apps that embed Studio surface keybinding information programmatically.

— Describes capability but no exact method or endpoint named. v0.9.2

05 Pre-build lighter copies of heavy clips

NEW

40

Pre-build a lighter copy of a heavy clip ahead of time to reduce rendering overhead when the clip is first used.

— States benefit but no command or config to invoke it. v0.9.2

06 **Producer reports browser-computed reference backgrounds** NEW

30

The producer now reports browser-computed reference backgrounds during a capture session.

— Thin description with no mechanism or usage detail. v0.9.5

└──▷ **BREAKING ON UPGRADE**

! `hyperframes cloud render` no longer accepts `--page-motion-blur`; use `data-motion-blur="on"` in the composition instead. Local `hyperframes render --page-motion-blur` is unchanged.

WAS THIS USEFUL?  

CopilotKit

Feedback ⓘ v1.84.0 NOTES ↗ RANK

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

> CopilotKit improved its generative UI prompt quality using a new cross-model evaluation suite, while introducing a breaking change that locks generative UI to canonical A2UI version 1.0.

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

01 Canonical A2UI 1.0 requirement for generative UI

BREAKING

65

Generative UI now requires canonical A2UI version `'1.0'` and capability key `'1.0'`; pre-release v1.0 operations and capability envelopes are no longer accepted. Demos must update their payloads and align agent SDK and CopilotKit versions accordingly.

— Names exact version strings but no migration command. v1.84.0

THINNER COVERAGE BELOW

02 Evaluated generative UI prompt improvements

IMPROVED

28

CopilotKit's generative UI prompt is now measured and improved across models using a new evaluation suite, intended to improve generation quality consistency across different underlying models.

— Describes benefit but no mechanism or metrics given. v1.84.0

└──▷ **BREAKING ON UPGRADE**

! Gen UI now requires canonical A2UI version `'1.0'` and capability key `'1.0'`; pre-release v1.0 operations and capability envelopes are no longer accepted — demos must update their payloads and align agent SDK and CopilotKit versions.

WAS THIS USEFUL? 👍 👎

OpenClaw

Feedback ⓘ v2026.10.10 NOTES ↗ CODE ↗ RANK

OpenClaw is a personal AI assistant platform that connects messaging channels, tools, and model providers through a gateway.

> OpenClaw's biggest change this window is secret-redacted PostgreSQL Git backups, alongside new local model tooling, expanded Android/iOS node capabilities, and incremental additions across MCP, sessions, cloud environments and diagnostics.

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

01 **Secret-redacted PostgreSQL Git backups** BREAKING 90

Git backups now apply `--exclude-secrets` to PostgreSQL exports and redact bundled-plugin credentials, webhook tokens, OAuth credentials, signing keys, and credential blobs. Existing backup history is not rewritten by this redaction, so affected integrations must sign in or register again after restoring a redacted backup.

— Names exact flag and credential types, plus migration note. v2026.10.10

02 **Session targeting via sessions_send** NEW 65

Sessions can now target conversation threads through `sessions_send` and run in existing folders on paired devices.

— Named command gives a concrete starting point. v2026.10.10

03 **Managed llama.cpp local model setup** NEW 65

Local models now support managed llama.cpp OCR, vision, and embedding-only setup, with automatic first-run inference selection and provider wait-time reporting.

— Specific capabilities named, no exact config shown. v2026.10.10

THINNER COVERAGE BELOW

04 **Cloud environment management additions** NEW 55

Cloud environments now support verified Node toolchains, repository recipes, saved profiles, startup progress, repository grouping, and local-to-cloud session moves.

— Six named cloud features grouped, no UI path given. v2026.10.10

05 **Expanded Android and iOS node capabilities** NEW 50

Android nodes now support system, photos, contacts, calendar, and motion capabilities; iOS nodes gain an authenticated background presence beacon.

— Lists capability names but no setup detail. v2026.10.10

06 **Plugin control over live run context** NEW 45

Plugins can now fork live context for background side questions, steer active runs, and override model fallbacks per run.

— Describes behaviour but no API or flag named. v2026.10.10

07 **Diagnostics export for message and run spans** NEW 45

Diagnostics now export message and run span prompts and final answers, with GC metrics attributed by kind and worker.

— Names data exported but no export command or format. v2026.10.10

08 **Secrets UX reveal and scoped egress** NEW 45

Secrets UX now supports reveal-on-demand for secure-entry prompts and scoped credential egress for setup grants.

— Names the two behaviours, no config or flag given. v2026.10.10

09 **MCP plugin apps via principal-bound relay** NEW 35

MCP plugin apps can now run across conversations and workspace files through a principal-bound relay.

— Names the relay mechanism but no usage path. v2026.10.10

10 **MCP event-triggered automations** NEW 30

MCP events can now trigger automations; session hovercards expose attached schedules.

— Thin description, no concrete surface named. v2026.10.10

└──▷ **BREAKING ON UPGRADE**

! Existing Git backup history is not rewritten by the new `--exclude-secrets` redaction; affected integrations must sign in or register again after restoring a redacted backup.

WAS THIS USEFUL?  

Anthropic Claude Agent SDK Python

Feedback ⓘ

v0.2.177 NOTES ↗ RANK

The Claude Agent SDK for Python is a developer library for building Python applications that run Claude-powered agents with tools, file access, and autonomous multi-step task execution.

> Claude Agent SDK Python adds fail-safe control over hook callback failures, letting developers decide whether an erroring hook blocks or permits an action.

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

01 **on_failure control for HookMatcher hooks** NEW 87

`HookMatcher` now accepts an `on_failure` option set to `"block"` or `"continue"`: `"block"` denies the guarded action when a callback raises, returns `None`, or returns a CLI-rejected value, while `"continue"` (the default) treats the failure as no decision. If `on_failure="block"` is set but the connected CLI does not support it, the SDK now raises `CLIConnectionError`.

Enforce that any hook callback failure blocks the guarded tool action — useful when you need a safety-critical check to never silently pass on error.

python

```python
from claude_agent_sdk import ClaudeAgentOptions, ClaudeSDKClient,
HookMatcher

async def strict_bash_check(input_data, tool_use_id, context):
    # Raises if the approval service is unreachable
    response = await approval_service.check(input_data)
    return response

options = ClaudeAgentOptions(
    allowed_tools=["Bash"],
    hooks={
        "PreToolUse": [
            HookMatcher(matcher="Bash", hooks=
[strict_bash_check], on_failure="block"),
        ],
    }
)

async with ClaudeSDKClient(options=options) as client:
    await client.query("Run: ./deploy.sh")
    async for msg in client.receive_response():
        print(msg)
```

— Names exact flag values and error type with runnable example v0.2.177

WAS THIS USEFUL? 👍 👎

cotal

Feedback ⓘ v0.135.9 NOTES ↗ RANK

The open standard for agent coordination

> Cotal shipped non-destructive inbox reading, per-space unavailability markers for catalogs, durable channel-family subscriptions, one-shot agent tasks, configurable channel history retention, and a new endpoint-listing tool, alongside several breaking changes to `@cotal-ai/core` exports and subscription permission enforcement.

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

01 Per-space unavailability markers in catalogs NEW 95

Mark a space as temporarily unavailable in a catalog entry with `"unavailable": "<reason>"` instead of its registration; commands targeting other spaces in the catalog keep working, and `cotal sync` lists the space under `unavailable`. Commands that target an unavailable space now refuse with the `DEPENDENCY_UNAVAILABLE` error and the catalog's reason string, rather than failing the entire catalog.

When a dependency space is down for maintenance, mark it unavailable in the catalog so the rest of your spaces stay reachable and `cotal sync` clearly reports the reason.

json

```
{
  "spaces": [
    {
      "name": "payments-api",
      "unavailable": "Undergoing scheduled maintenance until
2025-06-01"
    },
    {
      "name": "notifications",
      "registration": "..."
    }
  ]
}
```

— Exact config key, command, and error code with a runnable example v0.135.1

02 Channel family subscriptions with durable catch-up NEW 83

Subscribe an agent to a channel family such as `team.>` in its `subscribe` config to get durable catch-up on missed posts; `cotal_join` validates the pattern against the read ACL, and `cotal_channels` lists the family with its stored-copy state.

Subscribe an agent to a whole channel family so it catches up on posts it missed while offline.

```yaml
subscribe:
  - team.>
```

— Names config key and two tools with a config example v0.135.0

---

03 **Per-space channel history retention policy** NEW 80

Set a per-space channel history retention policy with `cotal history limit` so posts are dropped after an operator-chosen age rather than only by the default 1000-post-per-sender rule.

Cap how long a space retains channel posts to 14 days so old messages are pruned automatically.

```
$ cotal history limit 14d
```

— Runnable command with a concrete example and default behaviour noted v0.135.0

---

04 **One-shot agent tasks via cotal_spawn** NEW 78

Start a one-shot helper for a single task with `cotal_spawn` using `once: true` and a `prompt`; the manager relays the prompt as the only turn, the deadline defaults to one hour, and `-1` sets none; `goal-result` reads how the task ended.

— Names fields and defaults but no runnable example v0.135.0

---

05 **Subscribe pattern validation against allowSubscribe** BREAKING 78

An agent definition whose `subscribe` names a pattern that no single `allowSubscribe` entry contains (such as `perm.>` under `perm.*`) is refused when loaded, and `cotal_persona` refuses a `subscribe` channel outside `allowSubscribe`, writing nothing the manager could not load back.

— Describes exact refusal mechanism and example pattern, no migration command v0.135.0

06 **Peek support in cotal_inbox** NEW 75

`cotal_inbox` now supports `peek`, letting an agent read connector-managed automatic messages still queued in Codex, Jcode, and OpenCode seats without clearing them — a DM that arrives mid-turn or a channel post during any turn is now visible before that turn ends.

— Names exact tool, flag, and three connectors; no runnable example given v0.135.9

07 **liveKvEntries and liveKvValues call signature change** BREAKING 65

`liveKvEntries` and `liveKvValues` now refuse a call that passes options in the filter slot followed by a second options argument; callers must combine the signal and callbacks into one options object.

— Names both functions and the required fix, no code sample v0.135.0

THINNER COVERAGE BELOW

08 **Legacy chat history exports removed from core** BREAKING 55

`@cotal-ai/core` no longer exports `LEGACY_CHAT_HISTORY`, `isLegacyChatHistory`, or `reconcileChatHistory`, and `setupSpaceStreams` now returns nothing.

— Names removed exports but gives no migration path v0.135.0

09 **New cotal_endpoints tool** NEW 40

List all endpoints an agent can call with the new `cotal_endpoints` tool.

— Bare mention of tool name, no mechanism or example v0.135.0

└──▷ **BREAKING ON UPGRADE**

! `@cotal-ai/core` no longer exports `LEGACY_CHAT_HISTORY`, `isLegacyChatHistory`, or `reconcileChatHistory`, and `setupSpaceStreams` now returns nothing.

! `liveKvEntries` and `liveKvValues` now refuse a call that passes options in the filter slot followed by a second options argument; callers must combine the signal and callbacks into one options object.

! An agent definition whose `subscribe` names a pattern that no single `allowSubscribe` entry contains (such as `perm.>` under `perm.*`) is refused when loaded.

! `cotal_persona` refuses a `subscribe` channel outside `allowSubscribe` and writes nothing the manager could not load back.

WAS THIS USEFUL?  

platypus

Feedback ⓘ v4.1.0 NOTES ↗ RANK

Self-hosted AI Agents for your whole team — on your infrastructure, your models, around the clock.

> Platypus's v4.0.0 is a major breaking release overhauling REST API conformance, deployment boot requirements, plugin configuration, database integrity constraints, and tool naming, while removing Dashboards entirely. The following v4.1.0 adds transcript summary cost visibility and a reordered Trigger run detail view.

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

01 Stricter REST request and response conformance

BREAKING

95

Clients and integrations must now remove unknown fields from requests, send Chat turns to `POST …/chat/:chatId/turns` with a trigger and no `id` / `workspaceId` / `messages`, omit `organizationId` or `workspaceId` on Agent/Skill/MCP/Provider creates, send `modelIds` as objects rather than bare strings, use `PATCH` for Chat/Trigger/Kanban card and comment/Webhook/A2A endpoint updates, send a `workspaces` array when removing a member, use `POST …/mcps/:mcpId/test` for a stored MCP test, address attachment writes through the Workspace route, and send `searchSource` and `apiMode` on Provider writes.

— Enumerates every changed endpoint, method and field required. v4.0.0

02 Per-resource plugin config environment variables

BREAKING

95

Plugin configuration moves from a single combined environment variable to individual `PLATYPUS_PLUGIN_CONFIG_<NAME>` variables, where `<NAME>` is the plugin name without the `@platypus/` prefix, uppercased, with hyphens replaced by underscores (e.g. `PLATYPUS_PLUGIN_CONFIG_DOCKER`). The old `PLATYPUS_PLUGIN_CONFIG` variable is removed and the backend refuses to start while it is set; operators must move each key of its JSON object into its own per-plugin variable and unset `PLATYPUS_PLUGIN_CONFIG`.

— Exact variable naming rule and migration steps given. v4.0.0

03 Database migration enforces NOT NULL and scope constraints

90

BREAKING

The upgrade backfills and constrains existing data; restoring the previous image tag does not undo it, so operators must take a database backup first. A deployment whose `provider.modelIds` still holds strings or unresolved objects fails migration 0101 until migrations 0047 and 0091 have been applied, and a Trigger whose stored config cannot be repaired stops the migration and names its `id`. `invitation.token`, `chat.run_heartbeat_at`, `chat.tags`, the Agent id arrays, `workspace.max_daily_summaries`, and `user.banned` are now NOT NULL, and the database rejects a Provider, Agent, MCP, or Skill holding both or neither scope, or any value outside the Chat status, Trigger type, Trigger run status/event type, Organization gate, or member role sets.

— Names every constrained column and migration dependency. v4.0.0

04 **Deployment boot validation and safety defaults** BREAKING

90

Deployments must now explicitly set `DATABASE_URL`, `BETTER_AUTH_URL`, and `FRONTEND_URL` or the backend refuses to boot, naming the missing variable; there is no fallback to node-postgres' `PG*` variables or to localhost. `REQUIRE_INVITATION_TO_SIGN_UP` now defaults to `true`, so fresh deployments admit accounts only by invitation unless the operator sets it to `false`. With `NODE_ENV=production`, first boot refuses the well-known `admin@example.com` / `admin123` credentials and the shipped Compose `.example.env` leaves `ADMIN_PASSWORD` empty. An unsupported auth cookie topology, including a malformed `AUTH_COOKIE_DOMAIN`, is now a startup failure rather than a logged warning.

— Lists every env var and condition causing boot failure. v4.0.0

05 **Kanban and Triggers tools split into create/update pairs**

85

BREAKING

The Kanban and Triggers Tool sets now expose `createCard` / `updateCard`, `createComment` / `updateComment`, `createBoard` / `updateBoard`, `createColumn` / `updateColumn`, and `createTrigger` / `updateTrigger` in place of the former `upsertCard`, `upsertComment`, `upsertBoard`, `upsertColumn`, and `upsertTrigger` tools; Agents or Skills whose instructions name an old tool must be rewritten.

— Lists every renamed tool pair and required follow-up action. v4.0.0

06 **Plugin API v4 with stricter contract** BREAKING

80

Plugin API v4 now requires `ToolSetContext.timezone`, `ToolSetContext.sessionId`, and `ToolSetContribution.description` as non-optional fields, and `@platypuschat/plugin-sdk` major-bumps to 3.0.0. Plugin API v2

is dropped; a third-party Plugin declaring `apiVersion: 2` is refused at boot and must be moved to SDK v4 types.

— Names required fields and SDK version but no migration example. v4.0.0

07 **Docker sandbox compose file renamed** BREAKING 80

`compose.sandbox.yaml` is renamed to `compose.docker-sandbox.yaml`; operators must change `-f compose.sandbox.yaml` to `-f compose.docker-sandbox.yaml` in commands, scripts, and systemd units, remove `@platypus/docker` from `PLATYPUS_PLUGINS` in `.env` (boot fails on a Plugin listed twice), and move any Plugins added inside the old overlay into `.env`.

— Gives exact file rename and required config edits. v4.0.0

08 **Dashboards and widgets removed** BREAKING 70

Dashboards, widgets, and the Dashboards Tool set are removed with no export option; every Dashboard and its widgets are deleted on upgrade, and the `listDashboards`, `listWidgets`, and `updateWidgetData` tools are no longer available to any Agent.

— Names removed tools and data-loss consequence, no export path. v4.0.0

09 **Per-session Provider header substitution** NEW 65

Provider headers can now use the `${SESSION_ID}` placeholder in header values so each session is routed to isolated upstream resources.

— Names the placeholder but not where headers are configured. v4.0.0

THINNER COVERAGE BELOW

10 **Consistent scope and secret-presence fields in API responses** IMPROVED 55

Provider, MCP, Agent, and Skill API responses now always include `scope`, and `apiKeySet`, `headersSet`, and `bearerTokenSet` are always present; `oauthClientSecret` is never returned.

— Names exact fields but gives no endpoint paths. v4.0.0

11 **Transcript summary cost shown with task-model summaries** NEW 45

Transcript summaries are now written with the task model, and transcript views show the transcript summary cost in the divider.

— States the change but no cost format or location detail beyond UI. v4.1.0

12 **MCP OAuth re-authorization required** BREAKING 40

An MCP server authorized before Platypus recorded its OAuth authorization server can no longer refresh and must be signed in to again.

– States the consequence but no remediation steps. v4.0.0

13 **Batched tool-result clearing preserves prompt cache** IMPROVED

35

Tool results are now cleared in batches so the cached prompt prefix survives across long agent runs.

– Describes the change but no batch size or config. v4.0.0

14 **Agent turns answer at step ceiling** IMPROVED

30

Agent turns that hit the step ceiling now produce an answer rather than silently stopping.

– States the behaviour change but no mechanism or limit value. v4.0.0

15 **Response shown above Timeline on Trigger run view** IMPROVED

30

The Trigger run detail view now shows the Response above the Timeline.

– Bare UI reordering with no further detail. v4.1.0

└──▷ **BREAKING ON UPGRADE**

! Clients and integrations must remove unknown fields from requests, send Chat turns to `POST …/chat/:chatId/turns` with a trigger and no `id` / `workspaceId` / `messages`, omit `organizationId` or `workspaceId` on Agent/Skill/MCP/Provider creates, send `modelIds` as objects rather than bare strings, use `PATCH` for Chat/Trigger/Kanban card and comment/Webhook/A2A endpoint updates, send a `workspaces` array when removing a member, use `POST …/mcps/:mcpId/test` for a stored MCP test, address attachment writes through the Workspace route, and send `searchSource` and `apiMode` on Provider writes.

! A deployment whose `provider.modelIds` still holds strings or unresolved objects fails migration 0101 until migrations 0047 and 0091 have been applied.

! The upgrade backfills and constrains existing data; restoring the previous image tag does not undo it — take a database backup first. A Trigger whose stored config cannot be repaired stops the migration and names its `id`. `invitation.token`, `chat.run_heartbeat_at`, `chat.tags`, the Agent id arrays, `workspace.max_daily_summaries`, and `user.banned` are now NOT NULL, and the database rejects a Provider, Agent, MCP, or Skill holding both or neither scope, and any value outside the Chat status, Trigger type, Trigger run status/event type, Organization gate, or member role sets.

! `PLATYPUS_PLUGIN_CONFIG` is removed; the backend refuses to start while it is set. Move each key of its JSON object into its own `PLATYPUS_PLUGIN_CONFIG_<NAME>` variable, then unset `PLATYPUS_PLUGIN_CONFIG`.

! Deployments must now set `DATABASE_URL`, `BETTER_AUTH_URL`, and `FRONTEND_URL`; there is no fallback to node-postgres' `PG*` variables or to localhost.

- ! `REQUIRE_INVITATION_TO_SIGN_UP` now defaults to `true`; a fresh deployment admits accounts only by invitation unless the operator sets it to `false`.
- ! With `NODE_ENV=production`, first boot refuses the well-known `admin@example.com` / `admin123` credentials, and the shipped Compose `.example.env` leaves `ADMIN_PASSWORD` empty.
- ! An unsupported auth cookie topology, including a malformed `AUTH_COOKIE_DOMAIN`, is now a startup failure rather than a logged warning.
- ! Plugin API v2 is dropped; a third-party Plugin declaring `apiVersion: 2` is refused at boot and must move to SDK v4 types. `@platypuschat/plugin-sdk` major-bumps to 3.0.0.
- ! The Kanban and Triggers Tool sets no longer expose `upsertCard`, `upsertComment`, `upsertBoard`, `upsertColumn`, or `upsertTrigger`; they expose `createCard` / `updateCard`, `createComment` / `updateComment`, `createBoard` / `updateBoard`, `createColumn` / `updateColumn`, and `createTrigger` / `updateTrigger` instead — Agents or Skills whose instructions name an old tool must be rewritten.
- ! An MCP server authorized before Platypus recorded its OAuth authorization server can no longer refresh and must be signed in to again.
- ! Dashboards, widgets, and the Dashboards Tool set are removed with no export; every Dashboard and its widgets are deleted on upgrade, and `listDashboards`, `listWidgets`, and `updateWidgetData` are removed from every Agent.
- ! `compose.sandbox.yaml` is renamed to `compose.docker-sandbox.yaml`; change `-f compose.sandbox.yaml` to `-f compose.docker-sandbox.yaml` in commands, scripts, and systemd units, remove `@platypus/docker` from `PLATYPUS_PLUGINS` in `.env` (boot fails on a Plugin listed twice), and move any Plugins added inside the old overlay into `.env`.

WAS THIS USEFUL? 👍 👎

PydanticAI

Feedback ⓘ v2.65.0 NOTES ↗ RANK

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

> PydanticAI v2.65.0 adds a WebSocket transport for the OpenAI Responses API, a persistent ConversationStore with conflict detection, and refusal-aware model fallback, alongside several smaller model-specific additions and a breaking exception-type change for Z.ai quota errors.

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

01 ModelFallback on refusal

IMPROVED 75

`ModelFallback` now falls back to the next model when a model refuses, and `ContentFilterError` can be listed in `fallback_on` to match refusals.

Fall back to a second model automatically when the primary model refuses a prompt, by including ContentFilterError in fallback_on.

python

```python
from pydantic_ai import Agent
from pydantic_ai.models.fallback import ModelFallback
from pydantic_ai.exceptions import ContentFilterError

agent = Agent(
    ModelFallback(
        'openai:gpt-4o',
        'anthropic:claude-sonnet-4-6',
        fallback_on=[ContentFilterError],
    )
)
result = await agent.run('Sensitive prompt that may be refused')
print(result.output)
```

— Named exception and runnable code example provided. v2.65.0

02 ModelQuotaExceededError replaces rate-limit error for Z.ai

75

BREAKING

`ModelQuotaExceededError` is now raised instead of `ModelRateLimitError` for Z.ai balance and plan-limit 429 responses — code catching `ModelRateLimitError` for Z.ai quota errors will no longer trigger.

— Clear before/after and migration implication for exception handling. v2.65.0

03 **Allowed callers setting for Anthropic web tools** NEW

70

Adds `anthropic_web_tools_allowed_callers` setting to control `allowed_callers` on Anthropic's web tools, restricting which origins can invoke them.

Set allowed callers on Anthropic web tools to restrict which origins can invoke them.

python

```python
from pydantic_ai import Agent
from pydantic_ai.models.anthropic import AnthropicModel

agent = Agent(
    AnthropicModel(
        'claude-sonnet-4-6',
        anthropic_web_tools_allowed_callers=
['https://myapp.example.com'],
    )
)
```

— Named config field with a runnable code example. v2.65.0

04 **ConversationStore for persistent message logs** NEW

60

Adds `ConversationStore`, a new class that keeps a conversation as a log of messages stored once, and `Conversation.revision` to detect conflicting saves.

— Mechanism and class names given, no runnable example. v2.65.0

05 **Non-retryable model errors in Temporal config** NEW

60

Adds `non_retryable_model_errors` to Temporal configuration to stop retrying model errors that can never succeed on retry.

— Named config key, no example of its use. v2.65.0

THINNER COVERAGE BELOW

06 **12h prompt cache retention for GPT-5.6** NEW

50

Supports the `'12h'` prompt cache retention value on the OpenAI Responses API for GPT-5.6 and later.

— Names exact value and model constraint, no usage example. v2.65.0

07 **Clearer migration errors for removed V2 names** IMPROVED

45

Improves migration errors by naming the V2 replacement when code imports a removed name such as `OpenAIModel` or `MCPServerStdio`.

— Names examples of removed symbols, no runnable step. v2.65.0

08 **WebSocket transport for OpenAI Responses API** NEW 40

Adds WebSocket transport to the OpenAI Responses API, enabling real-time streaming over a persistent connection.

— Names the surface but gives no usage example or flag. v2.65.0

09 **Native mid-conversation system prompts** IMPROVED 35

Sends mid-conversation system prompts natively on OpenRouter OpenAI routes and the Claude models that support them.

— Describes behaviour but no mechanism or example. v2.65.0

└──▷ **BREAKING ON UPGRADE**

! `ModelQuotaExceededError` is now raised instead of `ModelRateLimitError` for Z.ai balance and plan-limit 429 responses — code catching `ModelRateLimitError` for Z.ai quota errors will no longer trigger.

WAS THIS USEFUL?  

Vercel AI SDK

Feedback ⓘ ai@7.0.162 NOTES ↗ RANK

The Vercel AI SDK provides TypeScript APIs for model generation, structured output, tool use, and streaming application interfaces.

> Vercel AI SDK added single-trace OpenTelemetry capture for durable WorkflowAgent runs and a steering.queue() method for scheduling follow-up messages after the current agent task completes.

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

01 **Single-trace telemetry for durable WorkflowAgent runs** NEW

91

A full durable `WorkflowAgent` run — prompt, final output, steps, and tools — can be traced as a single telemetry operation by passing the same `OpenTelemetryIdGenerator` instance as the tracer provider's `idGenerator` and `sampler` and as new OpenTelemetry({ durable }). Events are reported on completion or suspension with original timestamps, model spans are reported live with provider requests nested beneath them, and the enclosing `invoke_agent` tree is reconstructed when the run completes. Typed experimental durable context hooks are also available for telemetry integrations in `WorkflowAgent` runs.

Capture an entire multi-step durable WorkflowAgent run as one trace, so you can see all steps, tools, and the final output correlated under a single parent span in your observability platform.

typescript

```typescript
import { OpenTelemetry } from 'ai';
import { NodeSDK } from '@opentelemetry/sdk-node';
import { OpenTelemetryIdGenerator } from '@opentelemetry/sdk-
trace-base';

const idGenerator = new OpenTelemetryIdGenerator();

const sdk = new NodeSDK({
  idGenerator,
  sampler: myAlwaysOnSampler,
  traceExporter: myExporter,
});
sdk.start();

const telemetry = new OpenTelemetry({ durable: true, idGenerator
});

// Pass telemetry to your WorkflowAgent run so steps share one
trace
const result = await myWorkflowAgent.run(input, { telemetry });
```

— Names exact API, config, and behavior with runnable example `ai@7.0.162`

THINNER COVERAGE BELOW

02 **steering.queue() for post-task messages** NEW 50

Calling steering.queue() on an active agent run sends a message once the current agent task finishes.

— Names exact method but lacks usage example `ai@7.0.160`

WAS THIS USEFUL?  

AG-UI

Feedback ⓘ release/2026-11-03 NOTES ↗ RANK

AG-UI: the Agent-User Interaction Protocol. Bring Agents into Frontend Applications.

> AG-UI introduced a selectable A2UI Express inference path alongside the existing Direct JSON mode, and added structured error recovery for sub-agent generation failures.

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

01 A2UI Express inference path for adapters NEW

84

AG-UI's `ag_ui_adk` adds an A2UI Express inference path alongside Direct JSON, selectable per run for injected and explicitly owned render tools; Express output is compiled incrementally by `a2ui-agent-sdk` 0.10.0 into validated A2UI activity updates on a stable surface/activity identity. Inference mode is controlled per adapter with the new optional `inference_format` tool parameter, accepting `'direct_json'` or `'express'` (default `'direct_json'`), for adapters that support A2UI Express.

Switch an A2UI Express-capable adapter to Express inference mode instead of the default direct-JSON path.

python

```python
from ag_ui_a2ui_toolkit import run_a2ui_generation_with_recovery

result = await run_a2ui_generation_with_recovery(
    adapter=my_adapter,
    prompt=user_prompt,
    inference_format="express",
)
```

— Names module, SDK version and parameter values with a runnable example.

release/2026-11-02 release/2026-11-03

02 Structured error recovery for sub-agent generation NEW

81

Sub-agent errors can now be raised as structured exceptions with the new `A2UIGenerationError` exception class; `run_a2ui_generation_with_recovery` records the failed attempt and feeds those errors back to the model on the next retry, within the existing attempt limit, while transport failures and cancellations still propagate unchanged.

Signal a recoverable sub-agent failure so the orchestrator feeds the structured error back to the model on the next attempt rather than crashing the run.

```python
from ag_ui_a2ui_toolkit import A2UIGenerationError

def my_adapter_call(prompt):
    result = call_sub_agent(prompt)
    if result.has_validation_error():
        raise A2UIGenerationError(result.errors)
    return result
```

— Names exception class and function with clear retry mechanism and example.

release/2026-11-02

WAS THIS USEFUL? 👍 👎

anton

Feedback ⓘ v2.26.11.3.1 NOTES ↗ RANK

self-improving agent

> anton added per-person admission control for shared live-app sessions, a write-relay approval call, and expanded the cloud pod prompt with always-on org and team skills.

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

01 **Write relay approval via approve_held_write** NEW 55

Requests held by the write relay can now be approved using `approve_held_write`.

— Names exact call but no usage example given. v2.26.11.3.1

02 **Per-person admission control for live-app slots** NEW 40

Shared live-app slots now support per-person admission control, limiting who can join a running session.

— Describes mechanism but no config surface or command shown. v2.26.11.3.1

03 **Always-on skills in cloud pod prompt** IMPROVED 25

Always-on org and team skills are now included in the cloud pod prompt.

— Bare statement with no mechanism or surface named. v2.26.11.3.1

WAS THIS USEFUL? 👍 👎

agent-app-orchestrator

Feedback ⓘ v7.1.0 NOTES ↗

RANK

Agent orchestration service for the Agent Landing Zone, built on Microsoft Foundry Agent Service and Microsoft Agent Framework. Coordinates specialized agents for conversational experiences with grounded knowledge, NL2SQL and MCP tools.

> agent-app-orchestrator introduces a managed Foundry knowledge toolbox with configurable authorization delegation, shipped as a coordinated platform release with new rollback constraints.

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

01 Managed Foundry knowledge toolbox with delegated authorization

NEW 85

Use the `FoundryToolbox` client with request-local Foundry call context, letting Foundry own downstream knowledge-source delegation instead of a parallel Search-token path. Authorization mode can be set to `user_delegated` (default) or explicit hosted `service_identity`, letting sources enforce their own supported authorization without a source-type or capability allowlist. The toolbox is bound via the `TOOLBOX_ENDPOINT` environment variable as part of the coordinated platform release.

– Names client, both auth modes and the env var, but no usage example v7.1.0

THINNER COVERAGE BELOW

02 Coordinated redeployment required for auth-mode changes

45

BREAKING

Authentication-mode changes require coordinated redeployment and new sessions — rolling back a single component in isolation is not supported; operators must roll back the complete platform/component combination together.

– Clear operational constraint but no specific commands or components named v7.1.0

└──▷ **BREAKING ON UPGRADE**

! Authentication-mode changes require coordinated redeployment and new sessions — rolling back a single component in isolation is not supported; roll back the complete platform/component combination.

WAS THIS USEFUL? 👍 👎

beeai-framework

Feedback ⓘ typescript_v0.1.32 NOTES ↗

RANK

Build production-ready AI agents in both Python and Typescript.

> beeai-framework split out hard dependencies from its transformers extras and added resilience features like chat model fallback and per-attempt tool timeouts, alongside MiniMax regional support and the removal of GoogleSearchTool.

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

01 Transformers extras no longer bundle torch/peft/outlines

85

BREAKING

`beeai-framework[transformers]` and `beeai-framework[all]` no longer install `torch`, `peft`, or `outlines` automatically; users of `TransformersChatModel` must install those packages separately, e.g. `pip install "beeai-framework[transformers]" peft outlines` after installing the appropriate PyTorch build.

— Exact pip command and package names given python_v0.1.86

02 GoogleSearchTool removed

BREAKING

60

`GoogleSearchTool` has been removed; agents using it must switch to `DuckDuckGoSearchTool` or `WikipediaTool`.

— Clear migration targets named, but no code sample typescript_v0.1.32

THINNER COVERAGE BELOW

03 FallbackChatModel for backend resilience NEW

55

Add a `FallbackChatModel` to the backend that tries a sequence of chat models in order, falling back to the next when one fails.

— Class named, mechanism described, no usage example python_v0.1.86

04 Per-attempt tool run timeout NEW

50

Set a per-attempt timeout on individual tool runs via the tool run options to cap how long each retry attempt may take.

— Mechanism explained but no exact option name typescript_v0.1.32

05 MiniMax regional endpoints and default model NEW

50

Select MiniMax regional (CN) endpoints when constructing the MiniMax adapter for deployments routed to Chinese infrastructure; MiniMax text models now default to `M3.1 Flash` when no model is explicitly specified.

— Names region and model default but no exact constructor syntax `typescript_v0.1.32`

---

06 **Execution limits forwarded through serve adapters** IMPROVED

45

Forward execution limits through the serve adapters (`A2A` , `MCP`) so per-run caps are respected by hosted agents.

— Names adapters but no config or flag detail `python_v0.1.86`

└──▷ **BREAKING ON UPGRADE**

! `GoogleSearchTool` has been removed; agents using it must switch to `DuckDuckGoSearchTool` or `WikipediaTool` .

! `beeai-framework[transformers]` and `beeai-framework[all]` no longer install `torch` , `peft` , or `outlines` automatically; users of `TransformersChatModel` must install those packages separately, e.g. `pip install "beeai-framework[transformers]" peft outlines` after installing the appropriate PyTorch build.

WAS THIS USEFUL? 👍 👎

Nous Research Hermes

Feedback ⓘ v0.22.6 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 v0.22.6 adds a migration path for users coming from other agents, per-gateway isolated terminal tabs, and per-skill write locks.

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

01 **hermes migrate command for agent switching** NEW 75

Run `hermes migrate` or `hermes migrate <agent>` to move Grok Bot and OpenBot users over by installing the importer from the plugin catalog.

— Exact command named, but migration mechanics unexplained v0.22.6

THINNER COVERAGE BELOW

02 **Per-gateway terminal tabs** IMPROVED 50

Each terminal tab now owns the gateway it was opened on, so terminals on different gateways are independent and switching gateways does not affect existing tabs.

— Explains behaviour change but no UI path or flag given v0.22.6

03 **Per-skill write locks** NEW 10

Hermes adds per-skill write locks.

— Named only in summary, no mechanism or usage detail v0.22.6

WAS THIS USEFUL? 👍 👎

yoagent

Feedback ⓘ v0.26.6 NOTES ↗ RANK

yoagent is a Rust agent loop that streams from 7 LLM protocols and runs tools until done, from a laptop to a Cloudflare Worker.

> yoagent shipped dynamic OAuth credential refresh with redaction of sensitive headers for MCP over HTTP, growth-based token accounting for execution limits, and a breaking rutis 0.20.0 upgrade that changes how plugins declare optional services.

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

01 OAuth credential refresh for MCP HTTP transport NEW

95

HttpTransport::with_authorization(Arc<dyn AuthorizationSource>) lets a token source supply credentials before every request without restarting the transport; on a 401 response, challenged(&AuthChallenge) is invoked to attempt one retry on the same MCP session, while a 403 is never retried. `McpError::Http` gained err.www_authenticate() to read the server's WWW-Authenticate header on 401 or 403 responses, available with or without an `AuthorizationSource`.

Use a short-lived OAuth access token with an MCP HTTP transport that refreshes credentials automatically on 401 without tearing down the session.

```rust
use std::sync::Arc;
use yoagent::mcp::{AuthChallenge, AuthorizationSource,
HttpTransport};

struct MyTokenSource { /* token store */ }

#[async_trait::async_trait]
impl AuthorizationSource for MyTokenSource {
    async fn authorization(&self) -> String {
        format!("Bearer {}", self.current_token().await)
    }
    async fn challenged(&self, _challenge: &AuthChallenge) ->
Option<String> {
        // Refresh and return new token, or None to give up
        let new_token = self.refresh_token().await.ok()?;
        Some(format!("Bearer {}", new_token))
    }
}

let transport = HttpTransport::new("https://mcp.example.com")
    .with_authorization(Arc::new(MyTokenSource { /* … */ }));
```

— Named API, full retry mechanism, and a runnable code example. v0.26.6

02 **Growth-based token accounting for execution limits** NEW

90

Passing `TokenAccounting::Growth` to ExecutionLimits::with_token_accounting() makes `max_total_tokens` track only the net token growth a run adds rather than every resent prompt: the conversation present at run start becomes the baseline, each turn contributes its growth delta, compaction shrinks count, and turns without usage count as zero. The existing `TokenAccounting::Processed` default is unchanged.

— Named config API and precise accounting mechanism, no code sample. v0.26.5

03 **Credential header redaction in MCP error text** NEW

65

Headers set via `with_header` — authorization, cookie, and any header name containing key, token, secret, auth, or session — are now redacted from quoted server error text, while User-Agent stays readable.

— Exact redaction rules named but no example to try. v0.26.6

04 **rutis 0.20.0 requirement and plugin service declarations**

65

BREAKING

yoagent-rutis 0.12.0 requires rutis 0.20.0: the npm, PyPI, and Go runtimes move to 0.20.0, and plugins must now declare services they can run without in `optional` and read them via `ctx.get`.

— Exact version and config fields named, no migration walkthrough. v0.26.5

NEW INTEGRATION GPT-6.1 Terra preset (Responses API)

NEW INTEGRATION GLM-5.3 FlashX preset

NEW INTEGRATION Qwen 3.8 Omni Flash preset

└──▷ BREAKING ON UPGRADE

! yoagent-rutis 0.12.0 requires rutis 0.20.0: the npm, PyPI, and Go runtimes move to 0.20.0, and plugins must now declare services they can run without in `optional` and read them via `ctx.get`.

v0.26.4 NOTES ↗ STABLE

yoagent adds `muse-spark-1.3` and `muse-wonder-1.0` price entries to `ModelConfig::meta` via prices::enable_bundled().

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.26.4 https://github.com/yologdev/yoagent.git
# already have the repo? check out this version:
$ git checkout v0.26.4
```

└──▷ BREAKING ON UPGRADE

! yoagent-rutis 0.11.0 requires moving the npm, PyPI, and Go runtimes to rutis 0.19.3; hosts must now start runtimes with rutis-host's `SpawnPlugin`; agent events no longer go on the rutis bus.

WAS THIS USEFUL? 👍 👎

Mastra

Feedback ⓘ @mastra/core@1.81.0 NOTES ↗ RANK

Mastra is a TypeScript framework for AI applications and agents, providing workflows, agent memory, RAG, tool calling, and evals.

> Mastra's core package adds experimental storage-backed MCP tasks for long-running tool, agent and workflow calls, deepens observability with span outcome tracking, model/provider analytics and saved view schemas, and adds authenticated channel attachment storage, browser cookie injection and factory-scoped sandbox env vars — alongside breaking removals in playground-ui and core subpaths and deprecated factory board APIs.

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

01 Model and provider dimensions for span queries NEW

85

aggregateSpans() can now group and count spans by `model` and `provider` using `groupBy` and `countDistinct.model` / `countDistinct.provider`, gated by the new `span-aggregate-model-dimensions` store feature; querySpans() can filter by `model` and `provider`, gated by the new `span-query-model-fields` store feature.

Filter provider spans in the observability store to analyse cost and performance for a specific model provider.

typescript

```
const { spans } = await mastraClient.querySpans({
  timeRange,
  where: { op: 'eq', left: { path: 'provider' }, right: {
literal: 'anthropic' } },
});
```

— Runnable query example plus named feature flags and fields @mastra/core@1.81.0

02 Authenticated channel attachment storage NEW

85

Channel agents now download authenticated binary attachments (ZIP, DOCX) into `channel-attachments/`, capped at 100 MiB by default; the cap can be controlled per agent channel with `maxAttachmentStorageBytes` in the `channels` config.

Let a channel agent save authenticated Slack attachments locally while keeping a 500 MiB cap instead of the default 100 MiB.

```typescript
new Agent({
  // ...
  channels: { adapters: { slack }, maxAttachmentStorageBytes: 500
* 1024 * 1024 },
});
```

— Named config key, default limit, and runnable override example `@mastra/core@1.81.0`

03 **Browser cookie injection via setCookies** NEW 85

`MastraBrowser` threads can have session cookies injected before or after the browser opens via the new setCookies() method, which overrides addCookiesToOpenBrowser() and addPendingCookies() for custom providers.

Sign a browser agent into a preview site behind your own login so the agent can access authenticated pages without the session token passing through the model.

```typescript
await browser.setCookies(
  [{ name: 'my_app_session', value: token, domain:
'app.example.com', path: '/' }],
  threadId
);
```

— Named methods with a runnable code example `@mastra/core@1.81.0`

04 **Span outcome and stop reason tracking** NEW 75

Custom spans can now record execution `outcome`, typed `stopReason`, and `stopReasonContext` via span.end(), letting traces distinguish failures, cancellations, limits, guardrail blocks, and observer interruptions.

— Names exact API and the distinct outcome states it enables `@mastra/core@1.81.0`

05 **Factory-scoped sandbox environment variables** NEW 70

Sandbox environment variables can be scoped to a specific factory by passing `env` as a function of sandbox context (carrying `factoryId` and `factoryName`) to `PlatformFactorySandbox` or Docker sandbox.

— Names surfaces but lacks a concrete usage example `@mastra/core@1.81.0`

06 **Factory board API deprecated** DEPRECATED 70

`@mastra/factory`'s defineBoard() and `MastraFactory` `boards` / `includeDefaultBoards` are deprecated; board fields `title`, `role`, and `next` must be migrated to `name`, `runner.context`, and `nextPhases` respectively.

– Gives exact old-to-new field mapping for migration `@mastra/core@1.81.0`

07 **Storage-backed MCP tasks** NEW 65

Mastra adds experimental storage-backed MCP tasks in `@mastra/mcp` so long-running tools, agents, and workflows return immediately with a task handle that can be polled, answered, or cancelled across instances and restarts.

– Explains mechanism but no concrete API call shown `@mastra/core@1.81.0`

08 **Saved view schemas for observability lists** NEW 60

Saved observability list views (traces, spans, threads, logs) can be validated with `savedViewStateSchema` or `savedViewStateEnvelopeSchema` imported from `@mastra/core/storage`, with filter shapes reusable via `filterNodeSchema`.

– Names exports but no usage walkthrough given `@mastra/core@1.81.0`

09 **Playground-ui component and prop removals** BREAKING 60

`@mastra/playground-ui` removes the `ChartCard` `description` and `summary` props along with several `MetricsCard.*` components and related table/legend APIs, and also removes `useFluidHover`, `FluidHoverHighlight`, and `CommandList.highlightClassName`.

– Lists removed APIs but no migration path given `@mastra/core@1.81.0`

THINNER COVERAGE BELOW

10 **Automatic route registration for SignalProvider** NEW 55

Adding getRoutes() to a `SignalProvider` automatically registers webhook and OAuth routes on `server.apiRoutes` without manual server config.

– Clear method name but no example of registration output `@mastra/core@1.81.0`

11 **inputAccepted field on OutputResult** NEW 50

An `inputAccepted` field on `OutputResult`, available in `processOutputResult`, preserves user messages from failed turns that accepted input before erroring.

– Named field but no usage pattern or trigger shown `@mastra/core@1.81.0`

12 **Core runtime subpath removals** BREAKING 45

`@mastra/core` removes the runtime-importable subpaths `@mastra/core/network/vNext` and `@mastra/core/telemetry/otel-vendor`.

– Names removed paths without replacement guidance `@mastra/core@1.81.0`

└──▷ BREAKING ON UPGRADE

! `@mastra/playground-ui` : `ChartCard` `description` and `summary` props removed, along with several `MetricsCard.*` components and related table/legend APIs.

! `@mastra/playground-ui` : `useFluidHover` , `FluidHoverHighlight` , and `CommandList.highlightClassName` removed.

! `@mastra/core` : Runtime-importable subpaths `@mastra/core/network/vNext` and `@mastra/core/telemetry/otel-vendor` removed.

! `@mastra/factory` : defineBoard() and `MastraFactory` `boards` / `includeDefaultBoards` are deprecated; board fields `title` , `role` , and `next` must be migrated to `name` , `runner.context` , and `nextPhases` respectively.

WAS THIS USEFUL? 👍 👎

claw-orchestrator

Feedback ⓘ v7.17.0 NOTES ↗ RANK

Run Claude Code, Codex, Antigravity, Grok Build and OpenCode behind one runtime, and a run only counts as done when checks it didn't write pass. Cross-engine fan-out, councils, session handoff, spend caps, MCP server, ACP agent.

> claw-orchestrator v7.17.0 adds an MCP tool to resume paused autoloops, a config key to restrict which tools agent and goal nodes can call, and two Codex reliability fixes around capacity errors and error surfacing.

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

01 disallowedTools restriction on agent and goal nodes NEW

90

Setting `disallowedTools` on `agent` and `goal` nodes restricts which tools that node may use; it is passed as `--disallowed-tools` to `claude` and `grok` agent nodes and `claude` goal nodes, and nodes running engines that would ignore the flag are refused at start.

Prevent a goal node from calling file-write or shell tools by declaring a disallowed-tools list in the node config.

yaml

```yaml
goal:
  disallowedTools:
    - write_file
    - bash
```

— Names the config key, the flag, affected engines, and includes a runnable config example.

v7.17.0

02 MCP tool to resume paused autoloops NEW

71

The new `autoloop_resume` tool lets the plugin or MCP resume a paused autoloop run, mirroring what the existing `POST /autoloop/<id>/resume` endpoint already allowed, with manager errors returned as-is.

— Names the tool and the endpoint it mirrors, but no sample invocation. v7.17.0

03 Codex capacity-retry backoff IMPROVED

60

A Codex goal blocked by a model-at-capacity error now resumes automatically on the same thread after a backoff configured within `capacityRetry`, instead of failing the node.

— Names the config key and new behaviour but no example of setting it. v7.17.0

THINNER COVERAGE BELOW

04 **Codex lastError now captures turn/completed failures** IMPROVED

48

A failed Codex turn's error now surfaces in `lastError` even when the app-server reports it only in `turn/completed`.

— States the bug fix precisely but offers no way to act on it directly. v7.17.0

WAS THIS USEFUL? 👍 👎

skardi

Feedback ⓘ v0.8.0 NOTES ↗

Skardi is a self-improving context framework for ai agents

skardi adds `SKARDI_CONFIG_DIR` env var override, `rss.deny_cidrs` egress policy, `login --control` / `mcp register` agent-onboarding commands, and a public `FeedFetcher` API.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch v0.8.0 https://github.com/SkardiLabs/skardi.git
# already have the repo? check out this version:
$ git checkout v0.8.0
```

Override the CLI config directory with the `SKARDI_CONFIG_DIR` environment variable, which the CLI reads and writes as `$SKARDI_CONFIG_DIR/config.yaml` in place of `~/.skardi/config.yaml`.

Inject an egress policy into a stock skardi-server from the context with the `rss.deny_cidrs` setting.

Register agents and sign in to a control plane with the new `login --control`, `mcp register`, and agent-onboarding CLI commands.

Adds a single-attempt `FeedFetcher` that surfaces `Retry-After` headers and exposes public fetcher types.

NEW INTEGRATION WhatsApp Cloud API webhook receiver

NEW INTEGRATION WhatsApp source pack over Open Connector

NEW INTEGRATION Lark source pack over Open Connector mirroring Feishu

NEW INTEGRATION Azure AI Foundry custom endpoint for LLM extract

NEW SUPPORTED PLATFORM Windows agent setup via install.ps1

WAS THIS USEFUL? 👍 👎

LangChain LangGraph

Feedback ⓘ CVE-2026-118667 CVE ↗

STABLE

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

`CVE-2026-118667`: Unencoded identifiers in LangGraph SDK v3 streaming can redirect authenticated GET requests (CVSS 4.3, affects `>= 0.4.0, < 0.4.6`)

▾ **1 more release in this issue** · 2026-09-28 → 2026-10-07

> CVE-2026-118668 CVE ↗ STABLE
>
> `CVE-2026-118668`: LangGraph JavaScript custom authentication may permit cross-user Store access (affects `>=0.1.2,<0.15.0`)

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Inference Providers & GPU Clouds

Baseten

Feedback ⓘ updated · Nov 3, 2026 NOTES ↗ RANK

Baseten deploys and serves machine learning models as autoscaling inference endpoints, with a model API catalog and dedicated GPU deployments.

> Baseten's latest changelog adds enterprise identity and auth (SSO/SCIM, OIDC, AWS AssumeRole), real-time inference via WebSocket and gRPC, expanded autoscaling and observability tooling, plus several breaking changes to API authentication and the Truss CLI.

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

01 Truss push flag removal and behavior changes BREAKING

65

The `--trusted` flag is deprecated and removed from `truss push`; separately, updates to `truss push` in version 0.14.0 change push behavior, so review the 0.14.0 release notes before upgrading.

– Names exact flag and command, points to release notes snapshot-20261103

02 OIDC and AWS AssumeRole authentication across workloads NEW

60

Adds `OIDC` and `AWS AssumeRole` authentication for training jobs, runtime workloads via Runtime OIDC, model deployments, and workspace integrations.

– Names specific auth mechanisms but no config steps given snapshot-20261103

03 Dashboard GPU and inference monitoring NEW

60

New dashboard monitoring capabilities include concurrent inference request monitoring, container restart tracking, GPU temperature and utilization metrics, per-replica metrics, and workspace-wide GPU usage.

– Lists concrete metrics but no navigation or API given snapshot-20261103

04 Log export, CLI filtering and streaming NEW

60

Adds log export to `OTLP` endpoints, CLI log filtering and streaming, downloading logs from the dashboard for offline analysis, and streaming logs from the terminal with the Baseten CLI.

– Names OTLP and CLI capabilities with clear usage paths snapshot-20261103

THINNER COVERAGE BELOW

05 Environment-scoped access and observability NEW

55

Adds environment-scoped API keys, environment-scoped logs and metrics, and environment-level access control to isolate and manage resources per environment.

— Describes scope and controls but no config path `snapshot-20261103`

06 **New REST API endpoints for billing and deployment control** NEW 55

Adds new REST API endpoints, including retrieving billing usage via API, terminating a deployment replica via API, and broader endpoints for deployment management.

— Named endpoint purposes given but no exact paths `snapshot-20261103`

07 **WebSocket and gRPC inference streaming** NEW 50

Baseten adds `WebSocket` and `gRPC` support for real-time model streaming and inference.

— Names protocols but no setup detail or endpoint `snapshot-20261103`

08 **Configurable autoscaling scale-down and utilization targets** IMPROVED 50

Improves autoscaling with a configurable scale-down rate and a target utilization setting to control replica scaling behavior.

— Names two settings but no exact config keys or values `snapshot-20261103`

09 **Truss 0.13.3 minimum version requirement** BREAKING 50

Action required: upgrade to Truss 0.13.3 — older versions are no longer supported.

— Exact version and action given, no further mechanism `snapshot-20261103`

10 **Autoscaling schedules for predictable traffic** NEW 40

Create autoscaling schedules to pre-scale capacity ahead of predictable traffic.

— Describes feature purpose but no schedule syntax `snapshot-20261103`

11 **SSO and SCIM for enterprise identity** NEW 35

Baseten now supports SSO and SCIM for enterprise identity management.

— Bare mention, no setup or scope detail `snapshot-20261103`

12 **Activity log filters** NEW 35

Filter activity log entries with activity log filters.

— Minimal description of filter mechanism `snapshot-20261103`

13 **API authentication error response format change** BREAKING

30

Upcoming change to API authentication error responses — existing integrations that parse authentication error response bodies may break.

— Warns of change but gives no new format snapshot-20261103

14 **API key format change** BREAKING

30

Upcoming change to Baseten API key format — existing keys will be affected; review the changelog notice and upgrade accordingly.

— No concrete new format or migration steps given snapshot-20261103

└──▷ **BREAKING ON UPGRADE**

! Upcoming change to API authentication error responses — existing integrations that parse authentication error response bodies may break.

! Upcoming change to Baseten API key format — existing keys will be affected; review the changelog notice and upgrade accordingly.

! The `--trusted` flag is deprecated and removed; remove it from any `truss push` invocations.

! Updates to `truss push` in 0.14.0 change push behavior; review the 0.14.0 release notes before upgrading.

! Action required: upgrade to Truss 0.13.3 — older versions are no longer supported.

WAS THIS USEFUL? 👍 👎

Nebius AI Cloud Feedback ⓘ

October 26 – November 1, 2026 (2026-11-02) BLOG ↗

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

Nebius AI Cloud adds DRA for GPU and RDMA workloads on Kubernetes 1.34+, generic SAML 2.0 SSO, and External Secrets Operator sync for SecretStash.

NEW SUPPORTED PLATFORM Managed Kubernetes now supports dynamic resource allocation (DRA) on Kubernetes 1.34 or later for NVIDIA GPU and RDMA/InfiniBand workloads.

NEW INTEGRATION Single sign-on now works with any SAML 2.0 identity provider, not only Okta, Microsoft Entra ID, and JumpCloud.

NEW INTEGRATION A new tutorial covers syncing SecretStash secrets to Managed Kubernetes using External Secrets Operator.

WAS THIS USEFUL? 👍 👎

AWS Amazon Bedrock Feedback ⓘ

New topic: Watermarking in Amazon Bedrock (2026-10-30) BLOG ↗ RANK

Amazon Bedrock provides managed foundation-model inference, customization, knowledge bases, and APIs for generative AI applications.

> Amazon Bedrock added documentation on watermarking, covering which models embed watermarks in their output and how to detect them.

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

01 **Watermarking support documentation** NEW 31

Amazon Bedrock now documents which models embed watermarks in their generated output and how to detect those watermarks.

– Describes capability but no model names, flags, or API details given.

New topic: Watermarking in Amazon Bedro…

WAS THIS USEFUL? 👍 👎

◆ Agentic Infrastructure

ArcBox

Feedback ⓘ v0.11.0 NOTES ↗ RANK

ArcBox runs containers, microVM agent sandboxes, and full Linux or macOS virtual machines on a Mac from one Rust daemon and CLI.

> ArcBox v0.11.0 adds QEMU build management with on-demand downloads and code signing, a new DatapathControl API for TCP manipulation, a vhost-user virtio-fs implementation, new virtualization backends, and renames the Fleet runner subsystem to ArcBox Runners.

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

01 **QEMU build management and on-demand download** NEW 85

ArcBox now pins the QEMU build version in `assets.lock` and installs it from a local archive via a new CLI import path. QEMU is downloaded on demand when the System VM falls back to software emulation, interrupted QEMU and asset downloads resume instead of restarting from scratch, and ArcBox requires its own code signature on the QEMU build before use.

— Names config file, CLI path, and signing requirement with clear mechanism v0.11.0

02 **vhost-user virtio-fs protocol implementation** NEW 85

ArcBox implements the vhost-user protocol including feature negotiation, `SET_MEM_TABLE` guest memory mapping, vring setup, and `SCM_RIGHTS` descriptor passing, with virtio-fs shares served through a new `FsServer`.

— Names protocol steps and server component with real mechanism v0.11.0

03 **DatapathControl API for TCP manipulation** NEW 80

A new `DatapathControl` API allows aborting TCP connections and injecting TCP resets, supports silent TCP handshake rejection, lets inspected relays be abandoned without sending a TCP close, and integrates IPv6 connection lifecycle with the silent TCP rejection behavior.

— Names API and several concrete TCP behaviors, lacks endpoint signature v0.11.0

THINNER COVERAGE BELOW

04 **Hardware virtualization status in SystemRuntimeStatus API** NEW 55

The `SystemRuntimeStatus` API now reports hardware virtualization status, the effective backend in use, and the state of on-demand component downloads.

— Names the API struct but no fields or usage shown v0.11.0

05 **Fleet runner subsystem renamed to ArcBox Runners** BREAKING

40

The Fleet runner subsystem is renamed to ArcBox Runners; any scripts, config, or tooling that references the Fleet name will need to be updated.

– Clear migration note but no further detail on scope v0.11.0

06 **New virtualization backends: Virtualization.framework adapter and HvMachine**

NEW

30

Adds a Virtualization.framework adapter (R4a) alongside a new `HvMachine` backend (R4b).

– Bare naming of two backends with no behavior described v0.11.0

└──▷ **BREAKING ON UPGRADE**

! The Fleet runner subsystem is renamed to ArcBox Runners; any scripts, config, or tooling that references the Fleet name will need to be updated.

WAS THIS USEFUL? 👍 👎

emisar

Feedback ⓘ v0.56.0 NOTES ↗ RANK

Emisar lets AI agents securely run infrastructure actions through MCP, with approval controls that protect production systems.

> Emisar's Nomad pack adds in-task network diagnostics, letting agents verify DNS and HTTPS connectivity from inside a running allocation.

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

01 **DNS and HTTPS probes for Nomad tasks** NEW

72

New `nomad.alloc_exec_node_dns` and `nomad.alloc_exec_node_https` probes, added in Nomad pack 0.4.17, run DNS and HTTPS checks from inside a selected Nomad task. The probes use fixed Node.js checks with bounded output and normal TLS verification.

– Names exact probe IDs and pack version, but no usage example. v0.56.0

WAS THIS USEFUL? 👍 👎

Tracer Cloud OpenSRE

Feedback ⓘ v0.1.2026.11.3 NOTES ↗

RANK

OpenSRE is an open-source toolkit for building AI SRE agents.

> OpenSRE's latest release adds GitHub sign-in over SSH and on remote servers for headless workflows, plus a shell UI indicator for new output while scrolled back.

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

01 **GitHub sign-in over SSH and remote servers** NEW 35

OpenSRE now supports signing in and connecting GitHub over SSH and on remote servers, enabling headless and server-based workflows.

— Describes capability but no exact command or flag given v0.1.2026.11.3

02 **New-output indicator in shell UI** IMPROVED 28

The shell UI now shows a new-output bar when the transcript is scrolled back, so activity is visible without losing your current position.

— Explains behavior but lacks UI path or config detail v0.1.2026.11.3

WAS THIS USEFUL? 👍 👎

Snowflake Cortex AI

Feedback

Nov 3, 2026: Cortex AI Gateway inference, observability, and cost management (General av

NOTES ↗ RANK

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

> Snowflake's Cortex AI Gateway reaches general availability, adding governed inference endpoints, per-request tracing, organization-wide spend reporting, budget/quota controls, OpenTelemetry integration, and a no-code analytics view.

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

01 **No-code analytics in gateway Observability page (preview)**

NEW 70

The Analytics tab on the gateway's Observability page in Snowsight (preview) charts model usage, request failures and latency, MCP tool calls, and conversation length from the trace table without writing SQL. In Snowsight, go to AI Gateway > Observability > Analytics tab, then select the metric and time range to chart.

Review trace analytics charts for MCP tool calls and conversation length without writing SQL.

> 📍 In Snowsight, go to AI Gateway › Observability › Analytics tab, then select the metric (model usage, request failures, latency, MCP tool calls, or conversation length) and time range to chart.

— Includes exact UI navigation path and metric list.

Nov 3, 2026: Cortex AI Gateway inferenc…

02 **Per-request tracing via AGENT_TRACE_TABLE** NEW 60

The gateway exposes a `AGENT_TRACE_TABLE` view for querying per-request observability data — models called, timing, token counts, and errors.

— Names the view but not its schema or query syntax.

Nov 3, 2026: Cortex AI Gateway inferenc…

03 **Gateway endpoint supports OpenAI and Anthropic API formats**

NEW 60

Inference requests can be sent to the gateway endpoint in OpenAI Chat Completions or Anthropic Messages API format, from official SDKs or coding agents.

— Names both API formats but omits the actual endpoint URL.

Nov 3, 2026: Cortex AI Gateway inferenc…

THINNER COVERAGE BELOW

04 **Org-wide AI spend reporting via AI_GATEWAY_USAGE_HISTORY** NEW

50

A new `AI_GATEWAY_USAGE_HISTORY` view lets teams report on AI spend across the organization.

— Names the view; no detail on columns or query examples.

Nov 3, 2026: Cortex AI Gateway inferenc…

05 **Budget and quota controls for AI spend** NEW

50

Track spend by team with shared resource budgets, and block individual users at a limit with per-user quotas.

— Describes mechanism but no config keys or limits given.

Nov 3, 2026: Cortex AI Gateway inferenc…

06 **OpenTelemetry trace export to gateway** NEW

40

Client applications can export OpenTelemetry traces to the gateway for unified request tracing.

— Brief mention; no setup steps or endpoint named.

Nov 3, 2026: Cortex AI Gateway inferenc…

WAS THIS USEFUL?  

Anthropic TypeScript SDK

Feedback ⓘ sdk-v1.4.0

NOTES ↗ RANK

The Anthropic TypeScript SDK provides typed JavaScript and TypeScript clients for Claude APIs, streaming, and tool use.

> Anthropic TypeScript SDK v1.4.0 adds new session error types for agent limit handling, a new field on External Keys and Workspaces, and GA API support for analytics, spend limits, and RBAC.

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

01 **GA analytics, spend limits, and RBAC API surface** NEW 48

Adds GA methods for analytics, spend limits, and RBAC (role-based access control) to the API surface, along with named beta types for unified metrics and RBAC delete responses, types for analytics `group_by` and spend limit period filters, and shared analytics order and increased request actor types.

— Lists several added type areas but no endpoints or flags sdk-v1.4.0

02 **Session limit error types for agents** NEW 45

Adds `tool_use_limit_reached_error` and `compaction_limit_reached_error` session error types for handling agent session limit conditions.

— Names two error types but no handling example sdk-v1.4.0

03 **user_profile_id field on Keys and Workspaces** NEW 30

Adds a `user_profile_id` field to the External Keys and Workspaces types.

— Names the field but no usage detail sdk-v1.4.0

WAS THIS USEFUL? 👍 👎

Anthropic Python SDK

Feedback ⓘ v2.4.0 NOTES ↗ RANK

The Anthropic Python SDK provides typed Python clients for Claude APIs, streaming responses, and tool use.

> Anthropic Python SDK v2.4.0 adds new session error types to distinguish tool-use and compaction limit terminations, plus a batch of typed models covering analytics, RBAC, and admin fields.

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

01 Session error types for tool-use and compaction limits NEW

60

Adds `tool_use_limit_reached_error` and `compaction_limit_reached_error` session error types so code can distinguish when a session ends because a tool-use or compaction limit was hit.

— Names exact error types but no usage example shown v2.4.0

THINNER COVERAGE BELOW

02 Typed models for analytics, RBAC and admin fields NEW

38

Adds named beta types for unified metrics and RBAC delete responses, types for analytics `group_by` and spend limit period filters, a `user_profile_id` field to External Keys and Workspaces, and shared analytics order and additional request actor types.

— Lists several named type additions but no usage detail v2.4.0

WAS THIS USEFUL? 👍 👎

AWS Amazon Bedrock AgentCore

Feedback ⓘ

Runtime: Higher concurrent connection limit for V2 agents (2026-11-02) BLOG ↗

RANK

Amazon Bedrock AgentCore provides managed agent runtimes, memory, identity, gateways, observability, and evaluation services.

> Amazon Bedrock AgentCore boosted Runtime concurrency for V2 agents and added a new CloudWatch metric to isolate platform-side latency.

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

01 **OverheadLatency metric for V2 agents** NEW 70

A new `OverheadLatency` CloudWatch metric is emitted in the `AWS/Bedrock-AgentCore` namespace for V2 agents, measuring time from request receipt to the agent's first response byte, excluding the agent's own processing time, to isolate platform-side latency.

— Names exact metric and namespace, clear monitoring starting point.

Runtime: OverheadLatency metric for V2 …

THINNER COVERAGE BELOW

02 **Higher concurrent connection limit for V2 agents** IMPROVED 55

Amazon Bedrock AgentCore Runtime V2 agents now hold up to 250 concurrent connections per session with the default 2-vCPU allocation, up from 16, with no configuration change required.

— Concrete before/after numbers but no config action needed.

Runtime: Higher concurrent connection l…

WAS THIS USEFUL? 👍 👎

SAM (Agent Mesh)

Feedback ⓘ v0.2.0 NOTES ↗ RANK

SAM is a private network for AI agents that lets each node publish, discover, and call tools and models across machines under identity-bound, deny-by-default policy.

> SAM's v0.2.0 opens the mesh to the outside world with policy-gated HTTP Ingress for a2a, mcp, and inference services, adds a Go SDK for embedding mesh members in-process, and ships an MCP-registry-listed stdio server, alongside breaking changes to schema migration and proxy deployment.

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

01 External ingress for a2a, mcp, and inference services NEW

98

Adds Ingress support exposing `a2a`, `mcp`, and `inference` services to HTTP clients outside the mesh at `https://<router>/ingress/<peer_id>/<a2a|mcp|inference>/<name>[/<subpath>]`, enforcing the same identity and policy checks used inside the mesh. Access is controlled via a new `allowed_ingress` field on `PolicyRole` in mesh policy and an `ingress:` block (`auth_mode: edge|passthrough`, optional `allow` list) on a service in `agentmesh-node.yaml`. Unauthenticated requests get a 401 with `WWW-Authenticate` and RFC 9728 metadata at `/.well-known/oauth-protected-resource`, and the node sets `X-Mesh-Ingress`, `X-Mesh-Principal`, `X-Mesh-Role`, and `X-Mesh-Task` headers from the verified credential in edge mode. `GET /.well-known/agent-card.json` on an ingress URL returns a backend A2A agent card with `supportedInterfaces` set to the public URL and, in edge mode, the mesh OAuth 2.1 security schemes.

Expose an A2A service to external HTTP clients in edge auth mode by adding an ingress block to the node config.

yaml

```yaml
services:
  - name: my-agent
    a2a:
      url: http://localhost:8080
    ingress:
      auth_mode: edge
      allow:
        - member:user:alice@example.com
```

Start the router with an ingress listener and a public URL so external clients can reach mesh services.

```
$ agentmesh-router \
  --ingress-addr :8443 \
  --ingress-url https://router.example.com \
  --ingress-tls-cert-path /etc/tls/tls.crt \
  --ingress-tls-key-path /etc/tls/tls.key \
  --public-control-plane https://cp.example.com
```

— Full endpoint paths, config keys, headers, and auth flow are all named `v0.2.0`

02 **Go SDK for embedding mesh members** NEW 95

The new `github.com/google/agentmesh/sdk/meshnet` package lets a Go program embed a mesh member: `Server.Up` enrolls and attaches to routers, `ListenA2A` serves an A2A agent, `HTTPClient` and `Discover` call providers, `WhoIs` identifies callers, `AttenuateToken` appends a `tar_block`, and `MCPServer` / `InferenceHandler` expose tools and the `/v1` endpoint in process.

Embed a mesh member in a Go service to serve an A2A agent to the mesh and on router ingress.

go

```
import "github.com/google/agentmesh/sdk/meshnet"

srv := &meshnet.Server{}
if err := srv.Up(ctx); err != nil {
    log.Fatal(err)
}
ln, err := srv.ListenA2A(ctx, "my-agent")
if err != nil {
    log.Fatal(err)
}
http.Serve(ln, myAgentHandler)
```

— Package path, every method name, and a runnable code sample given `v0.2.0`

03 **External-url now required behind proxies** BREAKING 83

`agentmesh-one` behind a proxy (including Cloud Run) now requires `--external-url`; the public URL is no longer inferred from `Host`, `X-Forwarded-Host`, or `X-Forwarded-Proto` headers. A new `agentmesh_one_proxied_requests_without_external_url_total` metric, with a log entry on first occurrence, flags proxied requests missing the required flag.

— Exact flag, headers removed, and detection metric all named `v0.2.0`

04 **Signing key rotation interval controls** NEW 70

Use `--key-rotation-interval 0` to select the default 24-hour rotation interval, or `--signing-key-path` to disable automatic rotation.

— Exact flags and default behavior named, no further mechanism v0.2.0

---

05 **Control plane schema migration to v21 breaks downgrade** 70

BREAKING

Do not downgrade a control plane database migrated to schema version 21 back to v0.1.1: the old binary does not know the `ingress` resource type, and a policy save from it removes every `allowed_ingress` grant.

— Clear migration warning with exact schema version and consequence v0.2.0

---

06 **MCP stdio server via npm package** NEW 63

The `@agentmesh-p2p/sdk` npm package now ships an `agentmesh` MCP stdio server executable, configured through `AGENTMESH_*` environment variables, and listed on the MCP registry as `io.github.google/agentmesh`.

— Names package and env var prefix but no setup detail beyond that v0.2.0

---

07 **New Prometheus metrics for ingress, tokens, and provider reach**

NEW 63

Adds `agentmesh_router_ingress_requests_total{mode,outcome}` and `agentmesh_node_external_ingress_requests_total{mode,outcome}` for ingress traffic, `agentmesh_router_revoked_tokens` for revoked tokens, and `agentmesh_node_provider_reach_total{result="shared"}` / `result="rate_limited"` for concurrent provider handshakes.

— Exact metric names given but no usage guidance beyond that v0.2.0

THINNER COVERAGE BELOW

---

08 **PodDisruptionBudget for control plane Helm chart** NEW 40

The Helm chart now deploys the control plane with a `PodDisruptionBudget`.

— One-line addition with no further mechanism described v0.2.0

└──▷ **BREAKING ON UPGRADE**

! Do not downgrade a control plane database migrated to schema version 21 back to v0.1.1: the old binary does not know the `ingress` resource type and a policy save from it removes every `allowed_ingress` grant.

! `agentmesh-one` behind a proxy (including Cloud Run) now requires `--external-url`; the public URL is no longer taken from `Host`, `X-Forwarded-Host`, or `X-Forwarded-Proto` request headers.

WAS THIS USEFUL?  

MotleyAI SLayer

Feedback ⓘ v1.5.0 NOTES ↗ RANK

SLayer is an embeddable semantic layer that lets AI agents and people query warehouse data through shared metric definitions.

> SLayer's v1.5.0 release adds a new Ossie import pathway, CLI diagnostics, and query-output refinements, while tightening several CLI behaviors in breaking ways.

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

01 Ossie document import via slayer import-osi IMPROVED

88

`slayer import-osi` imports Ossie documents written to the current core spec, including the flat document shape, dialect selection, field visibility, `FILTER`, and dataset-scoped metrics; malformed models or relationships are skipped and reported instead of halting the import.

— Names exact spec elements and documents failure handling v1.5.0

02 CLI version and debug flags NEW

65

Print the installed version with `slayer --version`; add `--debug` to any command to see full tracebacks on failing commands.

Check which version of SLayer is installed before scripting or filing a bug report.

```
$ slayer --version
```

— Names both flags but little mechanism beyond their purpose v1.5.0

03 Query output formatting and dimension sorting IMPROVED

65

Table output now applies column and measure number formats (for example `$23.36K`), while JSON output stays raw; queries with dimensions and no explicit `order` now return rows sorted by those dimensions in the order listed, with an explicit `order` still taking precedence.

— Describes formatting and sorting precisely with example v1.5.0

04 datasources refresh-samples command moved BREAKING

65

`slayer search refresh-samples` has moved to `slayer datasources refresh-samples`; the old form now prints an error pointing to the new command.

— Names old and new command paths exactly `v1.5.0`

---

05 **Demo datasource ingestion flags removed** BREAKING 60

`slayer datasources create demo` now rejects flags that limit ingestion such as `--include`, because it always ingests every demo table.

— Names exact flag and new always-ingest behavior `v1.5.0`

THINNER COVERAGE BELOW

06 **Direct search without --question flag** IMPROVED 50

Run `slayer search <term>` directly without needing to pass `--question`.

— Simple command shortcut with little further explanation `v1.5.0`

---

07 **DuckDB 1.1 minimum requirement** BREAKING 35

SLayer now requires DuckDB 1.1 or later.

— States version requirement only, no migration detail `v1.5.0`

└──▷ **BREAKING ON UPGRADE**

! `slayer search refresh-samples` has moved to `slayer datasources refresh-samples`; the old form prints an error pointing to the new command.

! SLayer now requires DuckDB 1.1 or later.

! `slayer datasources create demo` now rejects flags that limit ingestion such as `--include`, because it always ingests every demo table.

WAS THIS USEFUL? 👍 👎

◆ Inference Serving & Optimization

Experiential Labs Experiential

Feedback ⓘ

v0.7.228 NOTES ↗ RANK

Experiential is the open source, zero markup gateway for BYOK, self-hosted and 1000+ marketplace models. It learns from your traffic to cut costs, recommend better models, and train a specialized model you own.

> Experiential v0.7.228 introduces fleet-wide lane limits that cap pace, burst and in-flight requests across the whole worker fleet, with configurable spill behavior when limits are hit.

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

01 **Fleet-wide lane limits for gateway dispatch** NEW 90

Experiential adds fleet-wide lane limits with pace, burst, and in-flight caps plus a spill policy via `GatewayRungDispatchPolicy.fleet`. A `FleetLaneGate` can be injected into the native control plane with NativeControlPlane(fleet_lane_gate=FleetLaneGate(policy=GatewayFleetLanePolicy(token_p burst=..., spill=...))) to enforce fleet lane limits after per-worker rung admission. Fleet shed now acts as a hard cap — overflowed requests move to the next sibling lane instead of being force-pushed back, and a raising gate admits on the per-worker policy.

Cap token throughput across the whole fleet and queue excess requests rather than refusing them outright.

python

```python
NativeControlPlane(
    fleet_lane_gate=FleetLaneGate(
        policy=GatewayFleetLanePolicy(
            token_pace=50000,
            burst=5000,
            spill="queue"
        )
    )
)
```

— Names exact classes, config fields and includes a runnable code example. v0.7.228

WAS THIS USEFUL? 👍 👎

Niko1221 Strata

Feedback ⓘ v0.2.10 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 the all-cards split engine on Linux, ISTA-DAS Coder model support, DeepSeek vision on Apple Silicon, and 128K context on 8 GB cards with new CLI and config controls.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.2.10 https://github.com/Niko1221/Strata.git
# already have the repo? check out this version:
$ git checkout v0.2.10
```

NEW SUPPORTED PLATFORM Turn the all-cards split engine on or off from the Terminal with `strata launch --split-engine off` or `--split-engine on`, set it in the run config with `"split_engine": "off"`, or override with the environment variable `STRATA_SPLIT_ENGINE=0`; on Linux with two NVIDIA or four R9700 cards running Qwen3.8, `strata launch` now starts the split engine automatically.

NEW INTEGRATION NEW SUPPORTED PLATFORM Launch the ISTA-DAS Coder (DeepSeek-V4.1-Flash with 192 of 384 experts per layer) with `strata launch --model ISTA-DAS-Coder`, or select it from the app's model list.

Control drafting per launch with `strata launch --drafting`, choosing Auto, MTP, DFlash2 or Off; DFlash2 is now the default drafter for Qwen3.8, replacing MTP, and Auto overrides any drafter choice saved in the model's file.

Force the DeepSeek vision plain kernel with `DS4_VISION_NO_MPP=1`, or test the tensor-operation kernel fallback with `DS4_VISION_FORCE_MPP_FAIL=1`.

UPDATED SUPPORTED PLATFORM Strata now starts 8 GB cards at 128K context (131,072 tokens) by default, stepping down the KV window, dropping the MTP layer, and moving the output head to pinned host memory as needed — the context and request count no longer shrink on low-memory discrete cards.

└──▷ BREAKING ON UPGRADE

! A drafting setting saved by an older version that named both drafters (`both` or `mtp+dflash2`) now counts as Auto rather than loading both.

! In the Terminal, Auto now overrides a drafter choice previously saved in the model's file.

! DFlash2 answers one request at a time; the app now loads the model for one request slot with DFlash2 on Auto, where 0.2.9 filled in up to 4 concurrent slots with MTP —

users who relied on multiple concurrent requests under Auto will need to choose MTP or set the concurrent-request count explicitly in Expert mode.

WAS THIS USEFUL? 👍 👎

SGLang

Feedback ⓘ v0.5.23 NOTES ↗ CODE ↗ RANK

SGLang runs large language models through an inference server with efficient structured generation.

> SGLang's v0.5.23 release adds Paged Experts for running oversized MoE models on a single consumer GPU, new low-latency megakernel/monokernel decode plugins, a quantized-MoE LoRA engine, Google TPU7x support, and broad performance and model-support gains across speculative decoding, caching, and ten new models.

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

01 Low-latency megakernel and monokernel decode plugins NEW

80

The opt-in NVIDIA megakernel (`sgl-nvidia-mega`) and AMD monokernel (`sgl-amd-mono`) plugins enable single-user decode up to 44% faster on GB300 and 75% faster on MI350X for DeepSeek-V4.1-Flash.

— Names exact plugin flags and hardware-specific gains. v0.5.23

02 Google TPU7x support via SGLang-TorchTPU NEW

75

SGLang now runs on Google TPU7x through the `SGLang-TorchTPU` PyTorch-native plugin, distributed via the `lmsysorg/sglang:v0.1.0-tpu` Docker image, with tensor and expert parallelism across hosts.

Pull the TPU7x Docker image to serve an LLM on Google Ironwood TPU hardware with SGLang-TorchTPU.

```
$ docker pull lmsysorg/sglang:v0.1.0-tpu
```

— Includes runnable Docker pull command and plugin name. v0.5.23

03 Prefill throughput and first-token latency gains IMPROVED

75

`DeepSeek-V4.1` on GB300 now delivers up to 1.5x more prefill throughput under load and about 28% lower first-token latency on 32K to 64K prompts, while MoE models such as `DeepSeek-V4.1-Flash` , `GLM-5.3-Flash` , and `gpt-oss` reach the first token 30% to 58% sooner on 8K prompts on H200.

— Gives concrete before/after numbers and named models/hardware. v0.5.23

04 Paged Experts for oversized MoE models NEW

60

The opt-in Paged Experts feature lets MoE models larger than GPU memory run on a single consumer GPU, delivering up to 30x faster performance than CPU offload.

— Names scope and 30x gain but no flag or command given. v0.5.23

THINNER COVERAGE BELOW

05 **HiCache L3 storage for MLA prefix caches** NEW 50

DeepSeek-style MLA model prefix caches can now be kept in HiCache file-based L3 storage when running decode context parallelism.

— Names storage layer and scope but no config key given. v0.5.23

06 **Speculative decoding determinism and extensibility** NEW 50

Speculative decoding gains support for seeding rejection-sampling draft proposals for deterministic inference, out-of-tree platform hooks so external platforms can supply their own speculative worker and EAGLE draft graph runners, and CUDA attention DP for DFlash and DFlash2 speculative decoding.

— Names three mechanisms but no flags or configs given. v0.5.23

07 **LoRA engine for quantized MoE models** NEW 40

A new opt-in LoRA engine serves LoRA adapters on FP8 and NVFP4 quantized MoE models.

— No flag or command given, thin description. v0.5.23

08 **Support for ten new models** NEW 40

SGLang now supports ten new models: Muse Spark Open, Beam-1.0, MiniMax-M3.1-Flash, MiniCPM-V 4.7, Cosmos3-Nano-Sim, FastH3 Omni, FLUX.3 Image Klein, HY-Nomas, and MOSS-Image.

— Lists all named models but no usage or config detail. v0.5.23

09 **Removed `--speculative-skip-dp-mlp-sync` flag** BREAKING 40

The `--speculative-skip-dp-mlp-sync` flag has been removed.

— Exact flag named but no migration guidance given. v0.5.23

10 **MSA-v1 sparse attention backend and NVFP4 KV pool** NEW 30

Adds the `MSA-v1` sparse attention backend and an `NVFP4` KV pool.

— Bare names, no mechanism or usage detail given. v0.5.23

└──▷ **BREAKING ON UPGRADE**

! The `--speculative-skip-dp-mlp-sync` flag has been removed.

WAS THIS USEFUL? 👍 👎

OpenNMT CTranslate2

Feedback ⓘ v4.8.5 NOTES ↗ RANK

CTranslate2 runs Transformer models with a C++ inference engine, quantization, and CPU and GPU optimizations.

> CTranslate2 v4.8.5 adds prompt-prefix caching for Drama 3 text-to-speech inference and native CPU support for reference-audio encoding used in voice cloning.

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

01 **Drama 3 prompt-prefix caching** NEW 54

CTranslate2 now caches Drama 3 prompt prefixes using LRU eviction, with cache controls and hit/miss diagnostics, to speed up repeated TTS inference.

— Mechanism described but no concrete flag or API name given v4.8.5

WAS THIS USEFUL? 👍 👎

NVIDIA Triton Inference Server

Feedback ⓘ v2.74.0

NOTES ↗ RANK

Triton Inference Server is an open-source platform for deploying and serving AI models at scale across CPUs and GPUs.

> Triton Inference Server v2.74.0 tightens input-size enforcement on generate requests, adds support for newer vLLM releases, improves request cancellation handling in the TensorRT backend, and introduces configurable S3 addressing style.

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

01 **`--http-max-input-size` enforced on nested generate parameters**

IMPROVED 65

The `--http-max-input-size` limit is now enforced across nested parameters of generate requests, not just top-level inputs, closing a gap where oversized nested fields could bypass the check.

— Names the exact flag and the behavior change. v2.74.0

THINNER COVERAGE BELOW

02 **vLLM >= 0.30 support in vLLM backend** IMPROVED 50

The vLLM backend now supports vLLM >= 0.30, with prompts rendered before generation and encoding.

— Names version and a behavior change but no config details. v2.74.0

03 **Earlier request cancellation checks in TensorRT backend**

IMPROVED 45

Request cancellation is now checked before dispatch to the backend and before batch assembly in the TensorRT backend, allowing cancelled requests to be dropped earlier in the pipeline.

— Describes mechanism but no config surface to act on. v2.74.0

04 **Configurable S3 addressing style** NEW 25

Triton adds support for configurable S3 addressing style.

— Only named in summary with no detail or flag given. v2.74.0

WAS THIS USEFUL?  

NVIDIA Model Optimizer

Feedback ⓘ 0.48.0 NOTES ↗

RANK

NVIDIA Model Optimizer provides quantization, pruning, distillation, and other model-compression techniques for deployment.

> NVIDIA Model Optimizer's 0.48.0 release adds six new GGML-compatible weight-only quantization formats, extends GPTQ and codebook search with importance-weighting, streams layerwise calibration exports to per-layer checkpoint shards, and adds indexer quantization for DeepSeek-V4 and GLM-5.3-Flash.

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

01 **New GGML weight-only quantization formats** NEW 90

Adds six GGML-compatible weight-only quantization formats: `q8_0` (32-value blocks, built-in `q8_0` PTQ recipe), `iq1_s` and `iq2_xs` (256-value blocks, built-in `iq1_s` and `iq2_xs` PTQ recipes; final weight dimension must be divisible by 256, and Megatron export requires tensor and pipeline parallel sizes of 1), `iq2_xxs` (CUDA encoder, `general/ptq` recipe, 2.0625 bits per weight), `iq2_s` (CUDA encoder, `general/ptq` recipe, 2.5625 bits per weight), and `iq1_m` (CUDA encoder, `general/ptq` recipe, 1.75 bits per weight). All IQ formats share the 256-value block constraint and export to unified HF and Megatron checkpoints.

— Names all six formats, recipes, bit rates and export constraints. 0.48.0

02 **Layerwise calibration export streaming and operator coverage**

IMPROVED 85

Layerwise calibration gains `layerwise.export_dir`, which writes each decoder layer to its own quantized checkpoint shard as it finishes, eliminating a separate export_hf_checkpoint() pass; combined with `layerwise.checkpoint_dir`, interrupted runs resume without redoing finished layers. Layerwise calibration also now supports quantizing and calibrating operators outside transformer layers, such as `lm_head`.

— Two named config keys with clear resume mechanism and scope. 0.48.0

03 **Sparse-attention indexer quantization for DeepSeek-V4 and GLM-5.3-Flash**

NEW 85

Adds fake quantization of the sparse-attention indexer key cache and query for DeepSeek-V4 and GLM-5.3-Flash through the new `indexer_k_quantizer` and `indexer_q_quantizer`; enable them by importing the

`configs/ptq/units/indexer_k_nvfp4` and `configs/ptq/units/indexer_q_nvfp4` units into a recipe.

— Names quantizers, config units and target models precisely. `0.48.0`

04 **GPTQ support for GGML quantization formats** IMPROVED

80

GPTQ (`algorithm: {method: gptq}`) now supports the GGML formats `iq1_s` , `iq1_m` , `iq2_xxs` , `iq2_xs` , `iq2_s` , and `q8_0` ; `block_size` must be a multiple of the target format's block size (256 for the IQ formats).

— Exact config key and numeric constraint given, mechanism thin. `0.48.0`

05 **Importance-weighted codebook search for IQ encoders** IMPROVED

78

The `iq1_s` , `iq1_m` , `iq2_xxs` , `iq2_xs` , and `iq2_s` encoders now accept an optional per-input-column `importance` vector to weight squared error during codebook search; GPTQ gains a new `importance_weighted` option that passes the square root of its Hessian diagonal as this vector.

— Config option and mechanism named, but no runnable example. `0.48.0`

THINNER COVERAGE BELOW

06 **Nemotron 3.5 Super VL PTQ recipe** NEW

20

Adds a PTQ recipe for Nemotron 3.5 Super VL, mentioned only in the release summary with no further detail in the changelog.

— Model and recipe type named but no further specifics given. `0.48.0`

07 **DFlash2 draft architecture** NEW

10

Adds the DFlash2 draft architecture, mentioned only in the release summary with no further detail in the changelog.

— Only a name is given; no mechanism or usage detail. `0.48.0`

WAS THIS USEFUL?  

LMCache

Feedback ⓘ CVE-2026-107204 CVE ↗

LMCache reuses and shares language-model key-value caches across requests and serving instances to reduce repeated prefill work.

`CVE-2026-107204`: Unauthenticated code execution through the `/run` _script endpoint in LMCache through 0.5.5 (CVSS 9.8, affects `< 0.5.6`)

WAS THIS USEFUL? 👍 👎

Etalab OpenGateLLM

Feedback ⓘ 0.9.1 NOTES ↗ RANK

An open-source, unified interface for running and managing self-hosted LLMs.

> OpenGateLLM's 0.9.1 release overhauls usage tracking with PostgreSQL and Langfuse-backed recording, tightens organization and key management, and removes the legacy QoS, Celery and per-user budget systems in a breaking cleanup.

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

01 **PostgreSQL and Langfuse usage recording** NEW 62

Adds `PostgresUsageRecorder` to persist inference usage to PostgreSQL, adds Langfuse as an alternative usage data source in place of the internal recorder, and traces model-forward requests through Langfuse following a chat completions refactor.

— Names the recorder class and Langfuse integration but no config steps 0.9.1

THINNER COVERAGE BELOW

02 **Playground usability and access improvements** IMPROVED 57

Authenticated users can now retrieve their own organization and use a new search bar to filter organizations by name in the playground UI; ProConnect 2FA can be forced in the playground via `oauth2-proxy` alpha config; and external sidebar links are flagged and moved to the bottom of the playground sidebar.

— Names config and UI elements but each addition stays brief 0.9.1

03 **Removal of user budget and @hooks decorator** BREAKING 46

User budget and the `@hooks` decorator have been removed — any code or configuration depending on per-user budget enforcement will break on upgrade.

— Names the decorator but gives no migration path 0.9.1

04 **vLLM token usage in rate limiting** IMPROVED 43

Reads vLLM token usage and discounts cached tokens when applying rate limits.

— Explains behaviour but no limit values or config shown 0.9.1

05 **Removal of QoS, Celery worker and timeseries metrics** BREAKING 43

The current QoS system, Celery worker, and timeseries metrics have been removed — any deployment relying on those components will break on upgrade.

— States removal and impact but no replacement guidance `0.9.1`

06 **Mandatory organization membership** BREAKING 42

Organization membership is now mandatory for all users; users not assigned to an organization will be rejected on upgrade.

— Clear before/after but no migration mechanism given `0.9.1`

07 **Key revocation and expired-key visibility** NEW 30

Adds key revocation and PATCH-based key management, and displays expired keys in the keys list so operators can see the full key lifecycle.

— No endpoint or flag named, mechanism unspecified `0.9.1`

08 **JSON API logs** IMPROVED 25

API logs are now emitted as JSON.

— Bare statement with no format fields or config detail `0.9.1`

└──▷ **BREAKING ON UPGRADE**

! The current QoS system, Celery worker, and timeseries metrics have been removed — any deployment relying on those components will break on upgrade.

! User budget and the `@hooks` decorator have been removed — any code or configuration depending on per-user budget enforcement will break on upgrade.

! Organization membership is now mandatory for all users — users not assigned to an organization will be rejected.

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

mlxtop

Feedback ⓘ v3.2.0 NOTES ↗ RANK

The mlxtop terminal UI monitors local LLMs on Apple Silicon, showing loaded models, memory use, GPU activity, and generation speed.

> mlxtop's docs this window cover new provider authentication environment variables for Strata/TensorFold servers and native live throughput gauges for Strata.

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

01 **Provider authentication and connection env vars** NEW INTEGRATION

NEW 85

Authenticate to a Strata or TensorFold server by passing the key via `MLXTOP_PROVIDER_API_KEY`; `MLXTOP_PROVIDER_PORT` sets a custom port, and `MLXTOP_PROVIDER_URL` sets the server root URL (without `/v1`).

Watch KV occupancy, prefix-cache reuse, and throughput on a local authenticated TensorFold server.

```
$ MLXTOP_PROVIDER=tensorfold \
MLXTOP_PROVIDER_API_KEY="$(cat "/path/to/api-key")" mlxtop
```

Watch live decode and prefill speeds on a Strata server running on a non-default port.

```
$ MLXTOP_PROVIDER=strata \
MLXTOP_PROVIDER_PORT=9090 \
MLXTOP_PROVIDER_API_KEY="$(cat "/path/to/api-key")" mlxtop
```

— Exact env vars with runnable example commands shown product docs

02 **Live decode/prefill gauges for Strata** NEW 75

Live decode and prefill speeds in Strata are driven from its native `strata:live_tok_s` and `strata:live_prefill_tok_s_mean` gauges; without them, completion counters provide averages that are excluded from live throughput charts and slowdown alarms.

Watch live decode and prefill speeds on a Strata server running on a non-default port.

```
$ MLXTOP_PROVIDER=strata \
MLXTOP_PROVIDER_PORT=9090 \
MLXTOP_PROVIDER_API_KEY="$(cat "/path/to/api-key")" mlxtop
```

— Names exact gauges and fallback behavior but no direct command `product docs`

UPDATED INTEGRATION Sushi / mlx-serve adapter updated to read sushi: gauge names

v3.2.0 NOTES ↗ **LATEST** STABLE

mlxtop adds per-category Journal recording controls, GPU load hysteresis thresholds, proportional Overview panel scaling, persistent post-prefill rate display and cache panel gap-fill.

└──▷ **GET THIS VERSION** ▾

```
$ git clone --branch v3.2.0 https://github.com/maximpri/mlxtop.git
# already have the repo? check out this version:
$ git checkout v3.2.0
```

Control which categories mlxtop writes to the Journal by setting `journal.prompts`, `journal.llm`, `journal.queue`, `journal.memory`, `journal.paging`, `journal.gpu` and `journal.thermal` in `~/.config/mlxtop/config.json` or the new Journal section in Settings.

GPU load thresholds now use hysteresis: the busy band activates at `gpu_warn_load` (default 75%) and holds until load drops below `gpu_warn_exit`, and saturation clears 10 points below `gpu_critical_load`, replacing the previous fixed 80% threshold.

Overview panels now scale proportionally with the terminal at every size, keeping host charts near two fifths of the height between compact and tall layouts instead of letting memory, compression and paging charts absorb all the space.

The prefill rate now persists after a request finishes, labeled `LAST PREFILL`, until a newer prefill replaces it, and a fresh dashboard seeds its first reading from the server's cumulative finished prefill counters.

The cache panel now uses the prefix reuse of still-running requests when no completed request supplies an interval rate, keeping a reading through long decodes.

WAS THIS USEFUL?  

Ollama

Feedback ⓘ v0.40.14 NOTES ↗ CODE ↗ RANK

Ollama downloads, serves, and chats with local language models through a CLI and REST API.

> Ollama's v0.40.14 release fixes sampler defaults on OpenAI-compatible endpoints to respect model-packaged values, preserves richer conversation history in `/v1/responses`, and speeds up top-k sampling on MLX.

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

01 Correct sampler defaults on OpenAI-compatible endpoints

IMPROVED 83

`/v1/chat/completions`, `/v1/completions`, and `/v1/responses` now leave `temperature`, `top_p`, and penalty fields unset when the client omits them, so models run with their packaged defaults instead of being forced to 1.0 — matching existing `/api/chat` behavior. For example, deepseek-r1 ships `temperature=0.6` and `top_p=0.95`, which now apply correctly when a client omits those fields.

Use an OpenAI SDK client against a model with custom sampler defaults (e.g. deepseek-r1 ships `temperature=0.6`, `top_p=0.95`) without having those defaults overridden — simply omit temperature and top_p and the model's packaged values apply.

```
$ curl http://localhost:11434/v1/chat/completions -H 'Content-Type: application/json' -d '{"model": "deepseek-r1", "messages": [{"role": "user", "content": "Explain quantum entanglement."}]}'
```

— Names exact endpoints and fields with before/after behavior v0.40.14

THINNER COVERAGE BELOW

02 Readable conversation history preserved in /v1/responses

IMPROVED 48

`/v1/responses` now preserves readable input content, including prior assistant refusals and inline images, instead of flattening agent message history.

— Names the endpoint and fix but no usage example v0.40.14

03 Faster MLX top-k sampling for large vocabularies IMPROVED

40

Top-k sampling on MLX is now faster for large vocabularies, using chunk-based sorting instead of full-vocabulary argpartition.

— Explains mechanism but gives no way to invoke it v0.40.14

WAS THIS USEFUL? 👍 👎

vMLX

Feedback ⓘ v1.6.83 NOTES ↗ RANK

The vMLX server runs compressed MLX models on Apple Silicon with disk caching, paged memory, continuous batching, and hybrid SSM scheduling.

> vMLX v1.6.83 adds MiniMax-M3.1 inference with DSpark speculative decoding, a set of fused MoE/attention kernel tuning flags, a new MSA selection benchmark, and panel controls for cache tiers.

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

01 DSpark speculative decoding controls for MiniMax-M3.1 NEW

90

vMLX adds DSpark speculative decoding for DeepSeek-V4.1/MiniMax-M3.1, toggled via the `--dspark on|off|auto` flag and a matching panel control. `VMLX_DSPARK_EXPLORE` (default off) re-measures stale DSpark planner positions, and `VMLX_DSPARK_SAMPLED=0` forces the text lane onto the plain, non-sampled path instead of DSpark sampling. Per-request speculative stats can be read from the `/health` endpoint after a generation.

Check whether the drafter is active and read per-request speculative stats from the health endpoint after a generation.

```
$ curl http://localhost:8000/health
```

— Names flags, env var and endpoint with usage example. v1.6.83

02 MiniMax-M3 decode and verify kernel tuning flags NEW

85

Environment-variable toggles tune MiniMax-M3.1 kernels: `VMLX_JANGT_FAST=1` enables fast JANGT decode kernels (default off); `VMLX_M3_FUSED_ATTN` enables the fused MiniMax-M3 attention half for decode/verify rows (default off, later turned on by default); `VMLX_M3_E1_BITS` overrides the E1 verify bit-width for the lm_head projection path; `VMLX_M3_DECODE_OVERLAP_DEFER` opts into deferred-submit decode overlap; and `VMLX_M3_EXACT_FLAT` enables exact multi-row E1 projections for verify (default off).

Force the DSpark text lane to the plain (non-sampled) path while also enabling the fused attention kernel, to isolate performance of the fused path.

```
$ VMLX_DSPARK_SAMPLED=0 VMLX_M3_FUSED_ATTN=1 vmlx serve <minimax-m3.1-path>
```

— Five named env vars with defaults and a run example. v1.6.83

THINNER COVERAGE BELOW

03 **MSA selection benchmark comparing score-sheet ops to kernel**

NEW

55

Running the benchmark with `--select` contrasts the score-sheet operations implementation of the MSA selection path against the kernel implementation.

— Named flag and comparison but no metrics or example. v1.6.83

04 **Cache tier selection in panel UI** NEW

50

The panel UI lets users switch between RAM only, SSD only, and RAM + SSD cache tiers.

— UI toggle named but no config key or command given. v1.6.83

WAS THIS USEFUL? 👍 👎

KoboldCpp

Feedback ⓘ v1.124.2 NOTES ↗ CODE ↗ RANK

KoboldCpp runs GGUF language models locally as a single-file application with a built-in web UI and API.

> KoboldCpp's v1.124.2 release ships new launch flags for context pooling, RPC caching, and MoE memory, alongside YuE2 music generation, new image models via sd.cpp, and AI Horde Chat Completions worker support.

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

01 New launch flags for context pooling, RPC caching, and MoE memory

NEW 90

Adds `--parallelrequests` (used alongside `--parallelrequests`; defaults to `contextsize` if unset) to control the total KV context pool shared across concurrent requests, `--rpccache` to cache RPC host tensors to disk, and `--lazy-mode` plus `--moe-cache-mib` to expose upstream lazy loading and MoE cache size.

Run large batched inference sessions where individual requests are big — raise the shared KV pool so requests do not queue unnecessarily.

```
$ ./koboldcpp-linux-x64 --model /path/to/model.gguf --parallelrequests 4 --parallelrequests 32768
```

Speed up repeated RPC inference by caching host-side tensors to disk, avoiding retransfer on each run.

```
$ ./koboldcpp-linux-x64 --model /path/to/model.gguf --rpccache
```

Reduce VRAM pressure on large MoE models by capping the MoE expert cache and enabling lazy weight loading.

```
$ ./koboldcpp-linux-x64 --model /path/to/moe-model.gguf --lazy-mode --moe-cache-mib 2048
```

— Exact flags with runnable example commands for each v1.124.2

02 AI Horde Chat Completions worker support NEW 75

AI Horde now supports Chat Completions workers automatically on upgrade, enabling jinja, reasoning, tool calls, streaming, and vision over AI Horde; toggle chat completions mode at `https://lite.koboldai.net`.

— Named toggle URL and enabled capabilities v1.124.2

03 **YuE2 music generation with SheetSage2 transcription** NEW 70

KoboldCpp adds YuE2 music generation support with three ready-made templates `YuE2-No-LLM`, `YuE2-Small-LLM`, and `YuE2-Large-LLM`, plus SheetSage2 transcription support enabling conversion of songs into YuE2 ABC codes.

— Named templates and format but no usage example v1.124.2

04 **Transcribe tab in MusicUI** NEW 65

MusicUI now has a transcribe tab, accessible at `http://localhost:5001/musicui`, for use with Whisper models, useful for TTS models that require reference transcripts.

— Exact UI URL named, limited mechanism detail v1.124.2

THINNER COVERAGE BELOW

05 **MCP wizard and paste mode in KoboldCpp Agent** NEW 50

KoboldCpp Agent now includes a `/mcp` wizard for configuring custom MCP endpoints and a `/paste` mode for handling multiline text input.

— Named subcommands but no usage example given v1.124.2

06 **Animated WebP default for video preview** IMPROVED 50

Animated WebP is now the default preview format for video generation (GIF still supported via SDUI settings; AVI remains the default download format).

— Describes format change, no config path given v1.124.2

07 **/v1/responses API endpoint in Kobold Lite** NEW 50

Kobold Lite adds `/v1/responses` API endpoint support.

— Named endpoint but no usage detail beyond name v1.124.2

08 **New image generation models via sd.cpp sync** NEW 45

KoboldCpp adds support for three new image generation models via an sd.cpp sync: MOSS Image, Fibo 2 Edit, and Flux 3 Klein.

— Names three models, no further usage detail v1.124.2

WAS THIS USEFUL?  

Open WebUI Feedback ⓘ CVE-2026-109572 CVE ↗ STABLE

Open WebUI is a self-hosted chat interface for local and remote models, with RAG over uploaded documents, tool calling, and multi-user access control.

`CVE-2026-109572` : Terminal users can manage Terminals policies and other users' terminals via the terminal proxy (CVSS 7.6, affects `>= 0.8.6, < 0.11.4`)

▾ **15 more releases in this issue** · 2026-09-27 → 2026-09-28

> CVE-2026-109292 CVE ↗ STABLE
>
> `CVE-2026-109292` : Other users' session tokens disclosed via OAuth-mode connections, tool servers and terminals (CVSS 6.8, affects `>= 0.6.27, < 0.11.4`)

> CVE-2026-109297 CVE ↗ STABLE
>
> `CVE-2026-109297` : Instance-wide stall via automation recurrence rules that evade sub-daily frequency detection (CVSS 6.5, affects `>= 0.11.0, < 0.11.4`)

> CVE-2026-109288 CVE ↗ STABLE
>
> `CVE-2026-109288` : Viewer session token theft via javascript: URLs in chat message links (CVSS 6.4, affects `< 0.11.4`)

> CVE-2026-109294 CVE ↗ STABLE
>
> `CVE-2026-109294` : Any authenticated user can exhaust server memory via a compressed remote image in a chat message (CVSS 6.5, affects `>= 0.6.42, < 0.11.4`)

> CVE-2026-109286 CVE ↗ STABLE
>
> `CVE-2026-109286` : Any website can steal a signed-in user's session token via the community stats message handler (CVSS 8.1, affects `>= 0.7.0, < 0.11.4`)

> CVE-2026-109287 CVE ↗ STABLE
>
> `CVE-2026-109287` : Directory structure of inaccessible knowledge bases disclosed via an unscoped directory id (CVSS 4.3, affects `>= 0.9.6, < 0.11.4`)

> CVE-2026-109296 CVE ↗ STABLE

`CVE-2026-109296` : Any authenticated user can reach disabled model provider backends via an explicit backend index (CVSS 4.3, affects `< 0.11.4`)

`CVE-2026-109298` CVE ↗ STABLE

`CVE-2026-109298` : Any authenticated user can exhaust worker memory via a nested-repetition knowledge-search pattern (CVSS 6.5, affects `>= 0.11.0, < 0.11.4`)

`CVE-2026-109295` CVE ↗ STABLE

`CVE-2026-109295` : Any authenticated user can repeatedly stall a worker via file-search regex on the event loop (CVSS 4.3, affects `>= 0.11.0, < 0.11.4`)

`CVE-2026-109290` CVE ↗ STABLE

`CVE-2026-109290` : Any authenticated user can steal another user's session token via script embedded in a DOCX preview (CVSS 8.7, affects `>= 0.11.1, <= 0.11.3`)

`CVE-2026-109285` CVE ↗ STABLE

`CVE-2026-109285` : Knowledge base collaborators keep read and delete access to removed files via stale file metadata (CVSS 5.0, affects `>= 0.6.0, < 0.11.4`)

`CVE-2026-109289` CVE ↗ STABLE

`CVE-2026-109289` : Deactivated users keep terminal command execution via an already open terminal WebSocket (CVSS 6.3, affects `>= 0.8.8, < 0.11.4`)

`CVE-2026-109284` CVE ↗ STABLE

`CVE-2026-109284` : Read-only shared folder members can still add chats to the folder by forking (CVSS 4.3, affects `>= 0.11.0, < 0.11.4`)

`CVE-2026-109283` CVE ↗ STABLE

`CVE-2026-109283` : Revoked users keep signing in via token exchange when roles are only in the ID token (CVSS 5.4, affects `>= 0.11.1, < 0.11.4`)

`CVE-2026-109293` CVE ↗ STABLE

`CVE-2026-109293` : Tool source code disclosed to read-only users via the tool export endpoint (CVSS 6.5, affects `>= 0.11.0, < 0.11.4`)

WAS THIS USEFUL? 👍 👎

◆ AI / LLM Gateways

cc-router

Feedback ⓘ v6.6.0 NOTES ↗

本地运行的大模型聚合网关，GUI桌面端app，零代码部署，把Coding Plan、大模型 API 额度聚合成一个虚拟 Plan，一键接入 Claude Code、Claude Desktop App、OpenClaw、OpenCode 等工具。Bundle your scattered Token Plan, Coding Plan, and LLM API quotas into a single virtual Plan, and plug it into Claude Code, Claude Desktop App, OpenClaw, OpenCode, and more.

cc-router adds a step-by-step setup wizard and adjustable desktop UI zoom (80%–125%, default 90%) in Settings → General → Appearance.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v6.6.0 https://github.com/finch-xu/cc-router.git
# already have the repo? check out this version:
$ git checkout v6.6.0
```

Add a step-by-step setup wizard covering language and appearance, disclaimer, subscription, model selection, connection test, and AI tool connection, with a final setup summary.

Control desktop UI zoom by choosing 80%, 90%, 100%, 110%, or 125% in Settings → General → Appearance; changes apply immediately, persist across restarts, and default to 90%.

NEW INTEGRATION Cheaper Inference provider

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

zvec-grep

Feedback ⓘ v0.3.0 NOTES ↗ RANK

Searches a local workspace with combined keyword and semantic matching, from a CLI or over MCP.

> zvec-grep v0.3.0 rewrites its core in Rust for major speed and memory gains, simplifies the CLI around a new `zg` query syntax, adds four new agent integrations, and exposes a tunable for embedding concurrency during indexing.

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

01 Embedding concurrency control for indexing NEW 90

Control embedding concurrency while building or updating an index with `--index-embedding-concurrency <n>` (CLI) or `ZVEC_GREP_INDEX_EMBEDDING_CONCURRENCY` (env var); capped at 8 for llama.cpp and Transformers.js, uncapped for Potion/model2vec (default 2), and uncapped for remote models.

Throttle embedding concurrency to 1 to recover from CUDA out-of-memory errors during index builds.

```
$ export ZVEC_GREP_INDEX_EMBEDDING_CONCURRENCY=1
zg --index --mode direct
```

— Exact flag, env var, defaults and caps with a runnable example v0.3.0

02 Simplified `zg` query CLI IMPROVED 75

Search directly with `zg "<query>"` using a new rg-inspired query + options format, with `--index` to build an index and `--status` for management operations.

Run a semantic search immediately from the terminal using the new simplified CLI format.

```
$ zg "where are user permissions checked"
```

— Names exact flags and a runnable query command v0.3.0

03 Rust-rewritten core execution pipeline IMPROVED 73

The core execution pipeline is now written in Rust, cutting CLI startup from 153 ms to 27 ms (83% faster), indexing from 3 min 39 s to 1 min 18 s (65% faster), and peak memory from 4.11 GiB to 1.92 GiB (53% lower) on a 75,128-document benchmark.

— Concrete before/after numbers but no new user-facing command v0.3.0

04 **New agent integrations: Copilot, Grok Build, VS Code, Pi** NEW

65

Configure agent integrations interactively via `zg --install`, now including GitHub Copilot, Grok Build, VS Code, and Pi alongside existing clients.

Wire zg into all supported AI agent clients (including the newly added GitHub Copilot, Grok Build, VS Code, and Pi) in one step.

```
$ zg --install
```

— Names the integrations and the exact install command v0.3.0

WAS THIS USEFUL? 👍 👎

Docling

Feedback ⓘ v2.146.0 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 v2.146.0 release expands its VLM-based parsing pipeline with three new presets covering OCR, grounded extraction, and guarded multi-pass processing.

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

01 **New VLM presets: Unlimited-OCR 2, LightOnOCR-3, guarded Nemotron 3.5**

NEW

55

Adds the Unlimited-OCR 2 VLM preset and its layout parser for document conversion, `LightOnOCR-3` VLM presets with grounded output parsing for structured document extraction, and guarded Nemotron 3.5 multi-pass parsing to the VLM pipeline for more reliable page processing.

— Names three presets but gives no flag or config to invoke them v2.146.0

WAS THIS USEFUL? 👍 👎

infino

Feedback ⓘ v0.11.20 NOTES ↗ RANK

Embedded retrieval library built on Parquet. Fast, efficient, and scalable.

> infino's v0.11.20 release overhauls full-text ingest and search internals for speed, while also tightening catalog config handling and compaction merge efficiency.

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

01 **Full-text search latency cut up to 9.9x** IMPROVED 65

Full-text search now runs exact take and id reads, lazy takes on their own tasks, and breaks score ties by `_id`, cutting 100-byte query latency 1.8–9.9x.

— Concrete mechanism and measured improvement, but no actionable flag v0.11.20

02 **In-RAM shard building for full-text ingest** IMPROVED 60

Full-text ingest now builds shards in RAM, tokenizing rows as they arrive and linking sealed files into local storage, with the writer pool reserving two cores.

— Explains mechanism and resource allocation, no config surface given v0.11.20

03 **Compaction merge efficiency improvements** IMPROVED 60

Compaction merge now maps inputs in place, links outputs, and pools widths by documents, reducing merge time by roughly a third.

— Mechanism and measured improvement given, no runnable surface v0.11.20

THINNER COVERAGE BELOW

04 **Supertable config applied via catalog** IMPROVED 30

The catalog now applies the supertable config section to tables created or opened through it.

— Names config section but lacks mechanism detail or key path v0.11.20

WAS THIS USEFUL? 👍 👎

python-sdk

Feedback ⓘ v10.3.0 NOTES ↗ RANK

Official Python SDK for the Pinecone vector database

> Pinecone's Python SDK v10.3.0 adds per-query embedding task-type overrides and region-controlled processing for integrated-embedding indexes, alongside a breaking tightening of proxy URL validation for gRPC connections.

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

01 Per-query embedding task type override NEW

86

Override the embedding task type for a single search by passing `parameters={"task_type": "..."}` to EmbedQuery(field=..., query=..., parameters: dict[str, Any] | None = None), letting one field serve multiple query types without creating a second index or re-embedding data.

Run a fact-checking query against a field normally used for retrieval, without creating a second index.

python

```python
from pinecone import EmbedQuery, Pinecone

pc = Pinecone()
facts = pc.index("fact-checks")

results = facts.documents.search(
    namespace="claims-en",
    top_k=5,
    score_by=[
        EmbedQuery(
            field="body",
            query="Is the Eiffel Tower taller than 300 m?",
            parameters={"task_type": "fact_verification"},
        )
    ],
)
```

— Names exact class, parameter, and includes runnable example v10.3.0

02 Processing location control for embedded fields NEW

82

Set where text is processed at index creation time with FieldEmbedConfig(..., processing_location: str | None = None) (e.g. `"us"`), and read the configured value back via

`pc.indexes.describe(name).schema.fields["body"].embed.processing_location`.

Create an index that processes embeddings in a specific region to meet data-residency requirements.

```python
from pinecone import FieldEmbedConfig, Pinecone, SchemaBuilder

pc = Pinecone()
schema = (
    SchemaBuilder()
    .add_string_field(
        "body",
        embed=FieldEmbedConfig(model="text-embedding-3-large",
processing_location="us"),
    )
    .build()
)
pc.indexes.create(name="support-articles-us", schema=schema)
```

— Names config class, field path, and includes runnable example `v10.3.0`

---

03 **Stricter proxy URL validation for gRPC** BREAKING 80

GrpcIndex(..., proxy_url=...) and pc.index(..., grpc=True) now raise `PineconeValueError` at construction when `proxy_url` uses a scheme other than `http://` (e.g. `https://`, `socks5://`, `socks5h://`) or has no scheme; the fix is to change the scheme to `http://` for gRPC. A malformed `proxy_url` now raises `PineconeValueError` (a `ValueError` subclass) instead of `httpx.ValueError` or `httpx.InvalidURL`, so `except httpx.InvalidURL` no longer matches these errors.

— Clear migration note but no runnable example given `v10.3.0`

└──▷ **BREAKING ON UPGRADE**

! GrpcIndex(..., proxy_url=...) and pc.index(..., grpc=True) now raise `PineconeValueError` at construction when `proxy_url` uses a scheme other than `http://` (e.g. `https://`, `socks5://`, `socks5h://`) or has no scheme; change the scheme to `http://` for gRPC.

! A malformed `proxy_url` now raises `PineconeValueError` (a `ValueError` subclass) instead of `httpx.ValueError` or `httpx.InvalidURL`; `except httpx.InvalidURL` no longer matches these errors.

WAS THIS USEFUL? 👍 👎

knowhere

Feedback ⓘ v1.2.44 NOTES ↗ RANK

Knowhere extracts, parses, and outputs structured chunks ready for AI Agents and RAG.

> Knowhere improved document ingestion fidelity, changing how file names are stored and adding a repair mechanism for re-parsing DOCX files while preserving their leading content.

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

01 **DOCX re-parse via reparse flag** NEW 85

A new `reparse=true` option lets users force a re-parse of an already-ingested DOCX document when its content has changed, bypassing the duplicate-content rejection guard. Example payload: `{"document_id": "<existing_document_id>", "reparse": true}`.

Force a re-parse of an already-ingested DOCX when its content has changed, without triggering the duplicate-content rejection guard.

```json
{
  "document_id": "<existing_document_id>",
  "reparse": true
}
```

– Includes exact flag and runnable example payload v1.2.43

THINNER COVERAGE BELOW

02 **Explicit URL file_name on ingest** IMPROVED 50

When ingesting new documents, knowhere now stores the explicit URL `file_name` field instead of deriving the name from the URL basename (e.g. `original.md`). This applies to documents ingested after this change shipped.

– Names the field but no full ingest API shown v1.2.44

03 **DOCX title and intro paragraph preservation** IMPROVED 35

Parsing now preserves DOCX title and introduction paragraphs that appear before the first section heading, rather than dropping them.

– Describes behaviour change but no mechanism or scope detail v1.2.43

WAS THIS USEFUL? 👍 👎

rabitq-library

Feedback ⓘ v0.8.2 NOTES ↗ RANK

An official lightweight library for the RaBitQ algorithm and its applications in vector search.

> rabitq-library v0.8.2 simplifies index construction with a three-line workflow and speeds up incremental vector insertion into IVF indexes.

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

01 **Faster IvfIndex add path** IMPROVED 40

The `IvfIndex` add path is now faster, improving throughput when incrementally inserting vectors into an existing index.

— Names the IvfIndex component but no benchmarks v0.8.2

02 **Three-line index construction workflow** IMPROVED 30

A new workflow reduces the amount of code needed to construct an index to three lines.

— No code sample or API names given v0.8.2

WAS THIS USEFUL? 👍 👎

stashbase

Feedback ⓘ v2.18.6 NOTES ↗ RANK

stashbase is an open-source, local-first writing workspace where you write with Claude Code and Codex, drawing on your own sources.

> Stashbase's latest release focuses on document appearance controls and editor polish, adding granular text sizing and paired themes, in-place formula editing, and smoother HTML preview refreshes.

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

01 **Expanded Appearance menu for documents** NEW 69

The Appearance menu now lets you adjust document fonts, text size across seven steps, page colors, line width, and line spacing, with the Paper and Green themes each offering matching light and dark variants.

Switch to the Paper dark theme and bump text to a larger step to reduce eye strain during long review sessions.

📍 Open the Appearance menu, select Paper (Dark) under page color, then drag the text-size slider to the desired step (1–7).

— Names concrete controls and theme variants with a usage example. v2.18.6

THINNER COVERAGE BELOW

02 **In-place block formula editing and math parsing** IMPROVED 47

You can now edit block formulas by moving the caret into and out of their source, and single-dollar math now follows parsing rules automatically without needing a separate setting.

— Explains mechanism but no concrete setting name or UI path given. v2.18.6

03 **Smoother HTML preview refresh** IMPROVED 28

HTML previews now refresh without a blank reload and show a clearer completion indicator.

— Describes the change but no mechanism or exact UI detail. v2.18.6

WAS THIS USEFUL? 👍 👎

redis-vl-python

Feedback ⓘ v0.30.0 NOTES ↗ RANK

Redis Vector Library. The AI-native Python client for Redis.

> redis-vl-python added async support to SemanticRouter, let the MCP server run user-defined async Python functions as code tools, and sped up SemanticCache lookups with selective filters.

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

01 User-defined async code tools in MCP server NEW 91

Register your own async Python functions as MCP code tools in-process by decorating them with server.code_tool(...), or load them from external modules by listing the module under `server.code_tool_modules` and setting `server.allow_code_tools: true` in the MCP server config. Functions registered as code tools must accept a `ToolContext` as their first positional parameter; declare write access with `writes=True` and an optional `scope=` for per-tool auth checks.

Allow the `rvl mcp` server to import and serve the code tools defined in `mytools.py`.

yaml

```yaml
server:
  allow_code_tools: true
  code_tool_modules:
    - mytools
```

— Names exact config keys, decorator and parameter requirements v0.30.0

02 Async methods for SemanticRouter NEW 82

Adds async routing to `SemanticRouter` with new methods aroute(), aroute_many(), adisconnect(), and the `aindex` property; pass `redis_url` instead of a pre-built client when using SSL, unix-socket, Sentinel, or `RedisCluster` connections.

Route an incoming query asynchronously inside an async event loop — useful in FastAPI handlers or other async frameworks.

python

```python
from redisvl.router import SemanticRouter

router = SemanticRouter(name='my-router',
redis_url='redis://localhost:6379')
result = await router.aroute('How do I reset my password?')
print(result)
```

— Lists exact method names and a runnable usage example v0.30.0

THINNER COVERAGE BELOW

03 **Faster SemanticCache lookups with filters** IMPROVED

47

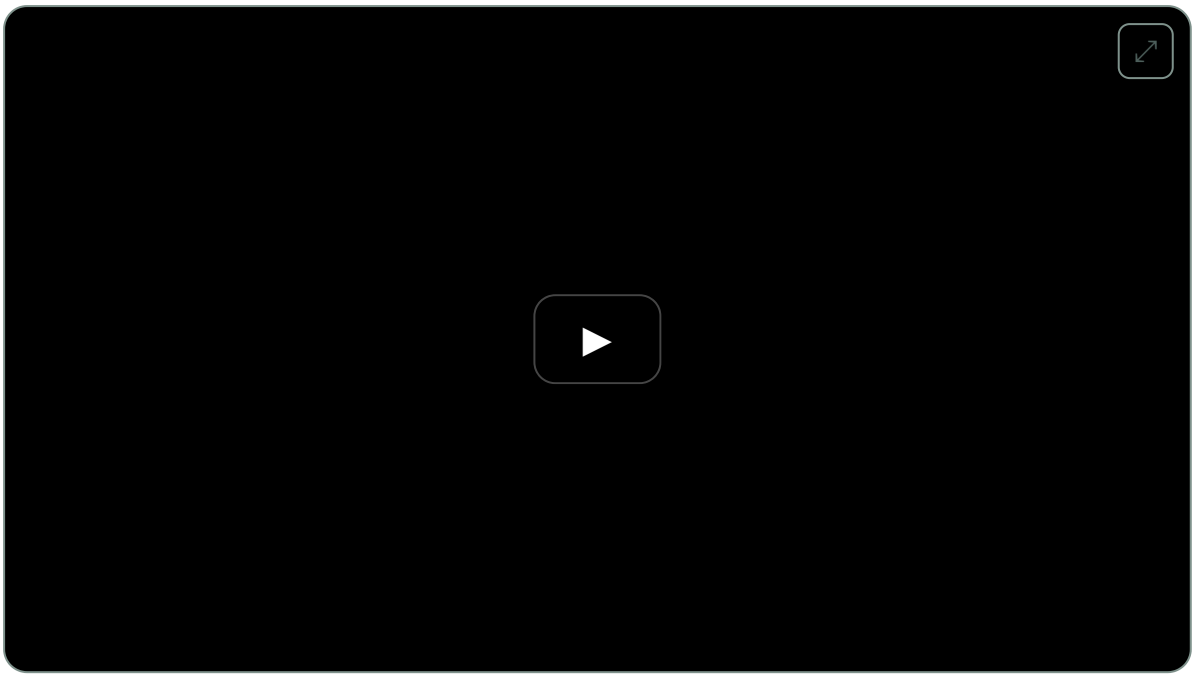
SemanticCache.check() and acheck() now run as KNN queries with the distance threshold applied client-side, cutting lookup time by up to a third when a selective filter expression is present.

— Explains mechanism and quantifies gain, no usage example given v0.30.0

WAS THIS USEFUL? 👍 👎

InfiniFlow RAGFlow Feedback ⓘ CVE-2026-119337 CVE ↗

RAGFlow is a retrieval-augmented generation engine built on deep document understanding, combining chunk-level citations with agentic workflows.



`CVE-2026-119337`: OpenAI-compatible agent completion bypasses private canvas authorization (CVSS 7.1, affects `>= 0.25.2, <= 0.27.2`)

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

Feedback ⓘ updated · Nov 3, 2026

NOTES ↗

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

LangSmith adds MCP tools for prompt management and sandbox file access, new deployment log fields, OAuth single-token support, and LSU-based Engine pricing.

Use `query_runs` in the LangSmith MCP to query trace runs with the same parameters as client.runs.query(), including typed selects for costs, token breakdowns, and full inputs and outputs, with cursor pagination; token counts and USD costs are returned by default.

Use `get_prompt_by_name` in the LangSmith MCP to retrieve the full prompt template for a named prompt.

New MCP tools let agents push prompt versions, browse commits, manage commit tags, update or delete prompts, and read and commit Context Hub agent and skill files.

The LangSmith MCP can now read, write, and search sandbox files, stream long-running sandbox commands, share sandbox ports behind LangSmith login, and report sandbox usage, organization usage, audit logs, usage limits, trace retention settings, and organization members.

Sandbox proxy rules now accept `match_headers` entries such as `'authorization: Bearer account-b'`, allowing several rules on one host to each swap a different placeholder for its own credential and serve multiple accounts.

+13 more

NEW INTEGRATION Vertex AI Model-as-a-Service open model cost resolution

NEW SUPPORTED PLATFORM GPT-6.1-Sol model support in Playground and model configurations

NEW SUPPORTED PLATFORM Claude Sonnet 5.5 support in Chat, Playground, Anthropic, Amazon Bedrock, and Vertex AI

NEW SUPPORTED PLATFORM Gemini 3.7 Flash and Gemini 3.8 Flash in Google Gemini and Vertex AI model selectors

UPDATED SUPPORTED PLATFORM OpenAI model cost resolution expanded to all pricing page models

WAS THIS USEFUL? 👍 👎

PromptLayer

Feedback ⓘ November 03, 2026 NOTES ↗ RANK

PromptLayer is a platform for managing, evaluating, and observing prompts and LLM applications.

> PromptLayer introduced a monthly spend cap for its provided provider keys with a self-serve fallback, added daily Vibe Chat and Wrangler message limits in Slack, and expanded request log tagging across several AI run types.

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

01 Monthly spend cap for PromptLayer-provided provider keys NEW

55

On Pro, Team, and Enterprise plans, a monthly spend cap now applies to PromptLayer-provided provider keys. Deployment runs are billed against the deployment allowance and do not count toward this dashboard spend cap, and when the cap is reached, runs prompt users to add their own provider API key instead of returning a generic error.

— Explains mechanism and plans but no dollar figure for the cap itself November 03, 2026

02 Daily message limits for Vibe Chat and Wrangler in Slack NEW

55

Vibe Chat now enforces a daily per-user message limit: paid plans get up to 100 messages per user per day, trial workspaces get up to 10, and users are notified in the app or Slack when the limit is reached.

— Exact limits given but no config to adjust them November 02, 2026

03 Request log tagging for Smart Table, assertions, simulator and signals

IMPROVED 50

Request logs now tag Smart Table AI data extraction, LLM assertion, and conversation simulator runs with their workspace, and tag signal classification runs by both signal and workspace, making all of these filterable in request logs.

— Names the run types and filter dimensions but not the UI steps November 02, 2026

WAS THIS USEFUL? 👍 👎

Arize Phoenix

Feedback ⓘ @arizeai/phoenix-client@7.18.0

NOTES ↗ RANK

Arize Phoenix is an open-source platform that monitors LLM application traces and evaluates their outputs.

> Arize Phoenix's client library adds typed resources for managing reusable dataset labels and assigning them to datasets.

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

01 Dataset label resources and assignment helpers NEW

70

The `@arizeai/phoenix-client/datasets` scope gains new typed resource and assignment helpers to manage reusable global dataset labels: list them with cursor pagination, and assign or replace labels on datasets selected by name or GlobalID.

– Names exact module and operations but no code sample @arizeai/phoenix-client@7.18.0

WAS THIS USEFUL? 👍 👎

Superlog Labs Superlog Feedback ⓘ

changes since 2026-10-04 CODE ↗

Superlog monitors AI agent telemetry with OpenTelemetry and groups signals into incidents for observability.

Superlog: New `PATCH /api/projects/:projectId/automation` endpoint sets `linearDefaultTeamId`, the Linear team for agent tickets; `null` or blank clears it.

└──▷ **TRY IT**

List the available Linear teams for a project so a user can pick one as the default for agent-filed tickets.

```
$ curl -X GET
https://api.superlog.io/api/projects/<projectId>/linear/teams \
  -H 'Authorization: Bearer <token>'
```

Set a specific Linear team as the default for agent tickets on a project, replacing the workspace fallback.

```
$ curl -X PATCH
https://api.superlog.io/api/projects/<projectId>/automation \
  -H 'Authorization: Bearer <token>' \
  -H 'Content-Type: application/json' \
  -d '{"linearDefaultTeamId": "<teamId>"}'
```

Slow insert pacing on a busy ClickHouse cluster with `insert_quorum` to grow batch sizes instead of flooding the replication queue, and spread connections evenly across replicas.

```
$ INGEST_CLICKHOUSE_MIN_INSERT_INTERVAL_MS=500
INGEST_CLICKHOUSE_KEEP_ALIVE=false node dist/ingest-consumer.js
```

Orgs can now set a monthly spend limit; once the estimated bill reaches the ceiling, billable features are frozen at their current overage while included allowances remain usable.

WAS THIS USEFUL? 👍 👎

MLflow Feedback ⓘ CVE-2026-116690 CVE ↗ STABLE

MLflow tracks, evaluates, and monitors machine learning and LLM applications, covering experiment tracking, tracing, prompt management, and a model registry.

`CVE-2026-116690`: Arbitrary file read on unauthenticated MLflow servers via prompt model version source (CVSS 7.5, affects `>= 3.10.0, < 3.17.0`)

▾ **2 more releases in this issue** · 2026-10-09 → 2026-10-13

> CVE-2026-116691 CVE ↗ STABLE
>
> `CVE-2026-116691`: MLflow CrewAI autologging records LLM provider API keys in logged trace spans (CVSS 6.5, affects `>= 2.19.0, < 3.17.0`)

> CVE-2026-116692 CVE ↗ STABLE
>
> `CVE-2026-116692`: Remote code execution on unauthenticated MLflow servers via third-party scorer deserialization importing caller-placed artifact modules (CVSS 9.8, affects `>= 3.12.0rc0, < 3.17.0`)

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI/LLM Security

containarium

Feedback ⓘ v0.126.0 NOTES ↗ RANK

Open-source agent runtime — SSH-native isolation, eBPF egress policy, Kubernetes + LXC backends, GPU passthrough, MCP-native CLI

> Containarium v0.126.0 adds a backup restore command and MCP tool, plus a new slides-agent recipe with a prebaked image for running presentation workloads in a sandbox.

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

01 **Backup restore command and MCP tool** NEW 75

Containarium adds a backup restore capability exposed both as a CLI command, `containarium backup restore --box <box-name>`, and as an MCP tool, letting users restore a previously taken box backup directly from the CLI without leaving the tool.

Restore a previously taken box backup directly from the CLI without leaving the tool.

```
$ containarium backup restore --box <box-name>
```

— Names exact command and flag; MCP tool mechanism unspecified v0.126.0

THINNER COVERAGE BELOW

02 **Slides-agent recipe with prebaked image** NEW 35

Containarium adds a slides-agent recipe with a prebaked open-slide image for running presentation workloads in a sandbox.

— Describes the recipe but no command or config shown v0.126.0

WAS THIS USEFUL? 👍 👎

cplt

Feedback ⓘ 2026.11.02-223908-6184a79 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 config key to scope GitHub Enterprise host restrictions within its gh command guard.

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

01 **Scoped GitHub Enterprise hosts in gh guard** NEW 75

The `gh_guard.hosts` config key in `.cplt.toml` lets users restrict the gh command guard to specific GitHub Enterprise hosts, so guard rules apply only to the listed instance (e.g. not to github.com by default).

Restrict the gh command guard to a specific GitHub Enterprise host so guard rules apply only to that instance and not to github.com.

```toml
gh_guard.hosts = ["github.example.corp"]
```

— Names exact config key and file with runnable example 2026.11.02-223908-6184a79

WAS THIS USEFUL? 👍 👎

ToolHive

Feedback ⓘ v0.60.0 NOTES ↗ RANK

ToolHive runs MCP servers in isolated containers and enforces identity and access policies for local and Kubernetes deployments.

> ToolHive's v0.60.0 release focuses on OAuth 2.0 token exchange hardening and scoping, adding Redis-backed grant storage, refresh token issuance, and a `--scopes` flag for LLM provider setup.

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

01 **Scoped LLM provider authentication** NEW 75

The `--scopes` flag can be passed to `thv llm setup` and `thv config set` to restrict the OAuth scopes requested when configuring an LLM provider, so the resulting token carries only the permissions the integration needs.

Restrict the OAuth scopes requested when connecting an LLM provider, so the token carries only the permissions the integration needs.

```
$ thv llm setup --scopes openid,profile,email
```

— Named flag with runnable example command v0.60.0

THINNER COVERAGE BELOW

02 **Delegate client token exchange hardening** IMPROVED 55

Several controls were added to the OAuth 2.0 delegate token exchange flow: DCR upstreams can now choose the token endpoint authentication method; `private_key_jwt` delegate clients now require an explicit trust mode before asymmetric-key token exchange is permitted; delegate token exchange is now limited to the configured source clients only, preventing unregistered callers from exchanging tokens; and the signing algorithm for delegate clients is now derived automatically instead of requiring manual configuration.

— Groups four thin security tweaks to same flow v0.60.0

03 **Redis-backed OAuth grant storage** NEW 45

OAuth 2.0 exchange grants can now be stored in Redis as an alternative to in-memory storage, enabling persistent grant state across restarts.

— Names mechanism but no config key given v0.60.0

04 **Refresh token issuance for token exchange** NEW 40

ToolHive now issues and rotates OAuth 2.0 refresh tokens through the token exchange flow, so delegate clients can maintain long-lived sessions without re-authentication.

— Explains behaviour but no config surface named `v0.60.0`

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Google Gemini API

Feedback ⓘ updated · Nov 3, 2026 NOTES ↗

RANK

Gemini API provides Gemini models for generating text, images, and other multimodal content in applications.

> Gemini API's biggest window update adds a beta `gemini-api` CLI, GA for `gemini-4-argon`, a new Voices endpoint with two TTS models and a 150+ voice library, plus new image (`gemini-nano-banana-2.1`), music (`lyria-3.5`) and Live API models — alongside major breaking changes: automatic routing off `gemini-2.5-pro` / `gemini-2.5-flash` after December 1, 2026, and a new Antigravity agent with renamed tool parameters.

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

01 Antigravity Agent 09-2026 with renamed tools and breaking parameter changes

90

BREAKING

Gemini API adds `antigravity-preview-09-2026`, a new Antigravity Agent version with new built-in tools including find_by_name(SearchDirectory, Pattern, MaxDepth) and grep_search(SearchPath, Query, IsRegex) for file and code search, replacing `antigravity-preview-05-2026`. In this version, built-in tool parameters switch from snake_case to PascalCase and file edits use line-range replacements (`replace_file_content`) instead of full rewrites; callers using `local_environment` or parsing `function_call` steps must update their tool invocations. `antigravity-preview-05-2026` shuts down October 5, 2026.

— Exact tool signatures, parameter changes, and shutdown date snapshot-20261103

02 Storage on by default for generation calls BREAKING

85

Gemini API now makes storage on by default (`store=true`) for new paid-tier projects using `generateContent`, `streamGenerateContent`, and `batchGenerateContent`; opt out per-request with `store=false` or via the Logs and Datasets dashboard Settings panel.

— Exact flags and dashboard path for opting out snapshot-20261103

03 Voices endpoint and TTS model family NEW

85

Gemini API adds the Voices endpoint (`/v1beta/voices`) alongside two new TTS models: `gemini-3.8-flash-tts` for studio-grade voice fidelity and `gemini-3.8-flash-lite-tts` for high-throughput production and real-time voice agent cascades. It also adds voice design, voice replication with consent verification, and an Extended Voice Library of 150+ prebuilt and custom voices for persistent custom vocal personas.

— Endpoint, two named models and library size given `snapshot-20261103`

04 **Automatic routing off Gemini 2.5 Pro and Flash after Dec 1 2026**

85

BREAKING

After December 1, 2026, API requests to `gemini-2.5-pro` automatically route to `gemini-3.8-flash` and are billed at Gemini 3.8 Flash rates, and requests to `gemini-2.5-flash` automatically route to `gemini-3.5-flash-lite` and are billed at Gemini 3.5 Flash-Lite rates. Both routings carry the same breaking API changes: function calling and `thought_signature` validation, `thinking_level` replacing `thinking_budget` , and removed sampling parameters and prefilled model turns.

— Exact replacement models, dates, and listed API-level changes `snapshot-20261103`

05 **Live API gains 3.8-live models with thinking levels** NEW

70

Gemini API adds `gemini-3.8-live` and `gemini-3.8-live-extended-thinking` to the Live API for real-time multimodal dialogue; the extended-thinking variant supports configurable thinking levels (`MINIMAL` , `LOW` , `MEDIUM` , `HIGH`) and asynchronous function calling.

— Named models and thinking-level values, no code sample `snapshot-20261103`

06 **Agentic video understanding across Flash models** NEW

70

Gemini API adds agentic video understanding for Gemini 3.7 Flash, 3.6 Flash, and 3.5 Flash-Lite across the Interactions and GenerateContent APIs, dynamically navigating video timelines on demand and using up to 88% fewer tokens for long-form content.

— Named models, APIs and token-reduction figure `snapshot-20261103`

07 **Deprecated models and agents** DEPRECATED

70

The `deep-research-pro-preview-12-2025` agent shuts down October 23, 2026 (update the `agent` parameter in `interactions.create` requests to `deep-research-preview-04-2026` or `deep-research-max-preview-04-2026`), and the `gemini-3.1-flash-image` model is deprecated in favor of migrating to `gemini-nano-banana-2.1` .

— Exact parameter and migration targets with shutdown date Clear migration target but no timeline or mechanism given

snapshot-20261103

08 **Nano Banana 2.1 image generation model** NEW 65

Gemini API adds `gemini-nano-banana-2.1`, a new image generation and conversational editing model supporting wide and panoramic aspect ratios (`1:4`, `4:1`, `1:8`, `8:1`) across `1K`, `2K`, and `4K` resolutions.

— Concrete aspect ratios and resolutions, no API call shown snapshot-20261103

09 **`gemini-api` CLI client** NEW 60

Gemini API adds a `gemini-api` CLI (beta), a stateless, zero-dependency command-line client for AI coding agents, scripts, and CI pipelines, supporting text and media generation, file analysis, and Gemini managed agents via the Interactions API.

— Names the CLI and its API but no commands shown snapshot-20261103

10 **Lyria 3.5 music generation model** NEW 60

Gemini API adds `lyria-3.5`, a new music generation model supporting text and image inputs and producing full-length songs as high-fidelity 44.1 kHz stereo audio.

— Specific audio format given, no invocation detail snapshot-20261103

THINNER COVERAGE BELOW

11 **Gemini 4 Argon model GA** NEW 35

Gemini API adds `gemini-4-argon`, a new model built for complex coding, agentic, math, and science tasks, now generally available.

— Just a model name and task list, no usage detail snapshot-20261103

└──▷ **BREAKING ON UPGRADE**

! After December 1, 2026, API requests to `gemini-2.5-pro` automatically route to `gemini-3.8-flash` and are billed at Gemini 3.8 Flash rates; breaking API changes include function calling and `thought_signature` validation, `thinking_level` replacing `thinking_budget`, and removed sampling parameters and prefilled model turns.

! After December 1, 2026, API requests to `gemini-2.5-flash` automatically route to `gemini-3.5-flash-lite` and are billed at Gemini 3.5 Flash-Lite rates; breaking API changes include function calling and `thought_signature` validation, `thinking_level` replacing `thinking_budget`, and removed sampling parameters and prefilled model turns.

! In `antigravity-preview-09-2026`, built-in tool parameters switch from snake_case to PascalCase and file edits use line-range replacements (`replace_file_content`) instead of full rewrites; callers using `local_environment` or parsing `function_call` steps must update their tool invocations.

! The `deep-research-pro-preview-12-2025` agent shuts down October 23, 2026; update the `agent` parameter in `interactions.create` requests to `deep-research-preview-04-2026` or `deep-research-max-preview-04-2026`.

! `antigravity-preview-05-2026` shuts down October 5, 2026; migrate to `antigravity-preview-09-2026`.

! The `gemini-3.1-flash-image` model is deprecated; migrate to `gemini-nano-banana-2.1`.

WAS THIS USEFUL? 👍 👎

Anthropic

Feedback ⓘ updated · Nov 3, 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 introduced server-side message threads (beta) for the Messages API, letting developers send only new messages per request to cut payload size and reduce time to first token in long conversations and agent loops.

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

01 Server-side message threads (beta) in Messages API NEW

91

Conversation history can now be kept on the server via message threads, enabled by including the `message-threads-2026-08-12` beta header; only new messages need to be sent in each request rather than the full history, reducing payload size and time to first token in long conversations and agent loops. Threads expire 5 minutes after the last request and are deleted no more than 1 hour after last use. The feature is available on the Claude API, Claude Platform on AWS, Claude Models on Google Cloud (Anthropic managed), and Claude in Microsoft Foundry.

— Names exact beta header, expiry/deletion timing, and supported platforms.

snapshot-20261103

WAS THIS USEFUL? 👍 👎

◆ Open-Weight Models

Z.ai GLM 5.3 FlashX Legal

Feedback ⓘ

released · Nov 3, 2026 MODEL ↗

GLM 5.3 FlashX Legal from Z.ai was released on 2 Nov 2026. 1049K-token context; $0.37 in / $1.25 out per million tokens (OpenRouter).

WAS THIS USEFUL? 👍 👎

DeepSeek V4.1 Flash

Feedback ⓘ price increase · Nov 2, 2026

MODEL ↗

DeepSeek V4.1 Flash is a multimodal model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4.1-flash-1017): $0.15 → $0.30 in, $0.60 → $1.20 out per million tokens (+100%), first seen 2 Nov 2026.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen3.8 27B

Feedback ⓘ context change · Nov 2, 2026

MODEL ↗ STABLE

Qwen3.8 27B is a multimodal model from Alibaba.

Context window grew from 1M to 1024K tokens on OpenRouter (qwen/qwen3.8-27b), first seen 2 Nov 2026.

▾ 1 more release in this issue · 2026-11-02

> price increase · Nov 2, 2026 MODEL ↗ STABLE
>
> Price increase on OpenRouter (qwen/qwen3.8-27b): $0.039 → $0.15 in, $1.25 → $1.875 out per million tokens (+285%), first seen 2 Nov 2026.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · November 3, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue