THE AI TOOLCHAIN

NO. 049

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED OCTOBER 7, 2026 · EVERY DAY

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

▸ // HOW THIS ISSUE IS MADE SHOW

VIEW

ISSUE VIEW

scroll



SCROLL | EXPAND IN PLACE | PREVIEW PANE

INFORMATION DENSITY

everything



BRIEFED | EVERYTHING

Every tool in the issue, in full.

all 252 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 46 tools

▾ // THE BRIEF HIDE

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

Three unrelated things stand out. Mistral put Large 4 (1M context, MoE, multimodal) into Public Preview and took OCR 4.1 to GA with block- word-level confidence scores you can threshold on. Inspect AI

vim ? help

now prices each request at the rate of the model that actually served it. And NVIDIA Model Optimizer's QAD workflow recovers accuracy lost to NVFP4 quantization.

BUILD

Extract typed fields from a document without writing a parsing pass, and know which OCR words to trust

Unstructured's new extraction API pulls JSON Schema-defined fields in one request or against an already-completed parse, so you stop hand-writing regex over element lists. Mistral OCR 4.1's block- and word-level confidence scores at GA let a pipeline route only low-confidence spans to human review instead of spot-checking whole documents.

Unstructured · Mistral AI

EVALUATE

Get a per-run cost number that survives a model fallback, and repeat samples without rewriting the task

Costs are now computed at the rate of the model that actually served the request, which matters when a gateway falls back or an alias resolves elsewhere — previously the eval report charged you at the model you asked for. `epoch=N` on enqueue_sample() repeats a sample in place, so variance estimates on a flaky task no longer require a bespoke loop.

Inspect AI

DEPLOY

Ship an NVFP4 model without eating the accuracy drop, and keep prefix reuse when you move to SGLang

Quantization-Aware Distillation trains the low-precision model against the full-precision teacher to claw back the quality 4-bit weights cost, and the Slurm launcher means the recovery run is a job submission rather than a hand-built pipeline across Hugging Face, Megatron-Bridge or Megatron-LM. LMCache's standalone SGLang integration with ZMQ/gRPC transport brings KV reuse to a second serving engine, so the choice of runtime stops dictating whether cached prefixes survive.

NVIDIA Model Optimizer · LMCache

GOVERN

Make a tool-calling agent refuse before it acts, and stop treating an unfenced login as acceptable

> Custom Guardrails on Agents/Conversations let you define the refusal policy yourself rather than inheriting the vendor's default, which is the gap most teams fill with a brittle pre-prompt classifier. Arcade now requires a passkey or second factor before the dashboard, CLI or MCP clients work at all — an agent framework holding third-party OAuth tokens is a credential store, and this closes the password-only path into it.
>
> Mistral AI

BUILD

◆ AI Coding Agents

Warp

Feedback ⓘ changelog · Oct 7, 2026 CHANGELOG ↗

Warp is an AI terminal that runs commands and helps developers build software.

Warp adds cloud agent handoff snapshots on interruption, fzf Alt-C shell support, a usage/cost popover, and team-switching UI improvements.

Adds `fzf Alt-C` support across bash, zsh, and fish shell integrations.

Adds a usage popover to the agent input footer showing conversation usage and cost.

Cloud agent runs now save a handoff snapshot when interrupted (SIGTERM/SIGINT) so in-progress work can be resumed instead of lost.

Adds the option to switch teams from Settings > Teams.

Moves the team switcher next to the profile avatar for easier access.

+3 more

NEW INTEGRATION AWS ECR credentials in managed secrets

NEW INTEGRATION ChatGPT sign-in for Warp agent inference

WAS THIS USEFUL? 👍 👎

Amazon Kiro

Feedback ⓘ changelog · Oct 7, 2026 CHANGELOG ↗

RANK

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

> Kiro's biggest update introduces reusable multi-agent Workflows spanning IDE, CLI, and Web, alongside new Power management commands, richer Steering context, slash-command prompts, and Claude Sonnet 5.5 availability.

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

01 Multi-agent Workflows across IDE, CLI, and Web NEW

95

Kiro adds reusable, multi-step multi-agent Workflows that run in the background while you keep working in the parent chat. The IDE exposes a `kiroAgent.workflows.enabled` setting; CLI V3 adds `/workflow run` and `/workflow` commands; Kiro Web adds `Settings > Workflows > Enable workflows`, an Upload workflow control accepting `.workflow.json`, `.workflow.yaml`, or `.workflow.yml` recipes, and Configuration Sync to import a full `.kiro/workflows/` folder. Three bundled recipes ship in Kiro Web: `investigate`, `feature-pipeline`, and `publish-pr`. A Workflows side panel adds Answer this step, Steer the agent, and Message this step's agent controls for per-step guidance plus pause, resume, stop, and retry-failed-steps actions, and a `Workflows: sub-agent tool` setting (toggleable under `/settings features`) controls whether the main chat can delegate directly to sub-agents or only through Workflows. Kiro also protects `.kiro/workflows/` recipes and agent/Hook/Power files by asking before any agent-initiated write even when broader file permissions allow it, and withholds untrusted-workspace recipes until the workspace is trusted.



Launch a reusable multi-agent Workflow from the CLI to deliver a feature end-to-end without leaving your terminal session.

```
$ /workflow run feature-pipeline
```

Enable Workflows in the IDE so multi-step background runs become available in the current workspace.

> 📍 In Settings, set `kiroAgent.workflows.enabled` to true, then open Agent Focus › Workspace Configuration and turn on Workflows, and start a new chat session.

— Names every setting, command, recipe, and protection mechanism.

changelog-20261007-250d6208

02 **Power install and uninstall commands** NEW 80

Adds `/powers install <name|path>` and `/powers uninstall <name>` slash commands in a local V3 session, plus `kiro-cli powers install <name|path>` and `kiro-cli powers uninstall <name>` for headless Power management.



Install a Power from the catalog inside an active V3 session so it is immediately available without restarting.

```
$ /powers install <name|path>
```

— Exact install/uninstall commands for both interactive and headless use.

changelog-20261007-250d6208

THINNER COVERAGE BELOW

03 **Folder and line-range references in Steering** IMPROVED

55

Adds `#[[folder:...]]` references and line-range selectors to Steering in CLI V3, enabling live one-level directory listings and partial file context that resolves when the Steering document loads.

— Describes syntax and resolution behavior but no worked example.

changelog-20261007-250d6208

04 **Configurable transcript scroll speed** NEW

55

Adds a Scroll speed control under `/settings display > Full Screen` to set transcript scroll to 1, 2, or 3 rows per mouse-wheel event (default 2).

— Exact setting path and configurable values given. changelog-20261007-250d6208

05 **Claude Sonnet 5.5 availability** NEW

55

Makes Claude Sonnet 5.5 available across Kiro for Pro, Pro+, Pro Max, and Power customers in FRA, with adaptive thinking on by default and a 1.3x credit multiplier.

— Names tiers and multiplier but no setup steps. changelog-20261007-250d6208

06 **Saved prompts as slash commands** NEW 45

Exposes workspace, global, and MCP saved prompts as slash commands in V3, with argument-hint completion and file-template or MCP content expansion before sending to the model.

— Explains behavior but no exact command syntax shown. changelog-20261007-250d6208

07 **SessionEnd Hook trigger** NEW 40

Adds a `SessionEnd` Hook trigger so session-level automation can run cleanup tasks reliably when a V3 session shuts down.

— Names the trigger but gives no usage example. changelog-20261007-250d6208

08 **V3 output style selection** NEW 20

Adds V3 output style selection in a V3 session, including styles supplied by upgraded agents.

— Bare mention with no detail on available styles. changelog-20261007-250d6208

WAS THIS USEFUL?  

Vercel v0

Feedback ⓘ snapshot-20261007 NOTES ↗ RANK

The v0 generative UI platform builds web applications from natural-language prompts.

> v0 converts personal accounts into collaborative team workspaces and adds permission controls for team skills and memories, alongside support for importing existing GitHub repositories as projects.

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

01 **Team permissions for skills and memories** NEW 65

Team members with edit access can now manage team skills and memories without requiring a Vercel write role, and a new Restrict Memories and Skills setting lets teams lock this down to admins only.

Lock down team-level memories and skills so only admins can modify them, preventing unintended changes by collaborators.

📍 In v0, go to Team Settings › Skills & Memories and enable the Restrict Memories and Skills toggle.

– Names exact setting and shows UI path to enable it snapshot-20261007

THINNER COVERAGE BELOW

02 **GitHub repository import as project** NEW 55

v0 now supports importing any GitHub repository into v0 as a project, letting users start generating and iterating on its UI with AI assistance.

Import an existing GitHub repo into v0 to start generating and iterating on its UI with AI assistance.

📍 In v0, go to New Project › Import GitHub Repository, select your repo, and confirm the import to create a GitHub-backed project.

– Exact UI navigation steps provided for import snapshot-20261007

03 **Team workspaces for collaboration** NEW 35

Personal v0 accounts can be converted into team workspaces, enabling users to invite others to collaborate on chats and projects.

– Describes the change but no setup steps given snapshot-20261007

WAS THIS USEFUL?  

Anysphere Cursor

Feedback ⓘ snapshot-20261007 NOTES ↗

RANK

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

> Cursor's latest docs update adds chat transcript search, finer-grained approval controls for MCP tools and protected file edits, and several agent workflow conveniences including PR context linking, remote file attachment, and conversation navigation shortcuts.

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

01 Attach local files to remote SSH/WSL agents NEW 75

Dragging a local file (e.g. `auth.log`) into the prompt box for an SSH or WSL agent uploads it to a separate uploads folder on the remote machine and references it, so the remote agent can read the file without polluting the workspace.

Attach a local log file to a cloud agent running over SSH so the remote agent can analyse it without polluting your workspace.

> 📍 In the Agents Window prompt for an SSH or WSL agent, drag your local file (e.g. auth.log) into the prompt box — Cursor uploads it to a separate uploads folder on the remote machine and references it so the agent can read it.

– Exact mechanism and a runnable drag-drop example product docs

02 Approval cards for confirmed MCP tools IMPROVED 60

MCP tools marked as requiring confirmation now always show an approval card in Auto-review, unless the tool is allowlisted or Run Everything is active.

– States exceptions and condition clearly product docs

03 Add pull request context via # picker NEW DETECTION NEW 60

Typing `#` in the Agents Window prompt and selecting a pull request from the picker that appears adds the PR as context to your message.

Quickly give an agent the context of an open pull request without leaving the prompt.

> 📍 In the Agents Window prompt, type `#` and select the pull request from the picker that appears — the PR is added as context to your message.

– Exact trigger character and steps given product docs

04 **Keyboard shortcuts for navigating agent messages** NEW

60

Press `Cmd+Up` / `Cmd+Down` to step between your own messages in a conversation, or `Cmd+Shift+Up` to jump to the top of the conversation.

Navigate a long agent conversation to review earlier decisions without scrolling.

> 📍 Press `Cmd+Up` / `Cmd+Down` to step between your own messages, or `Cmd+Shift+Up` to jump to the top of the conversation.

— Exact shortcuts for a runnable action `product docs`

THINNER COVERAGE BELOW

05 **Search within chat transcript** IMPROVED

57

Press `Cmd+F` to search the chat transcript whenever chat or prompt has focus. Matches are highlighted in yellow and the search no longer flickers to a false `0 of 0` state.

— Names exact shortcut and behaviour fix `product docs`

06 **Per-file toggle for Tab completions** NEW

50

Cursor Tab completions can now be turned on or off per file from the '...' menu on any code file tab.

— Names menu and scope, no screenshot or extra detail `product docs`

07 **Copy PR link from agent menu** NEW

45

An agent's Copy menu now includes a `Copy PR Link` option to copy the pull request link directly.

— Names menu item, no further mechanism given `product docs`

08 **Approval card detail for protected file edits** IMPROVED

45

Approval cards for out-of-workspace or protected config file edits now display the reason and full file path.

— Describes new fields shown, no workflow example `product docs`

09 **Named environments in Continue on Cloud** IMPROVED

35

Continue on Cloud options now display each saved environment's name, making it easier to distinguish between multiple environments on the same repo.

— Thin description of a UI label change `product docs`

`snapshot-20261007` NOTES ↗ LATEST STABLE

> Cursor adds iOS remote control for local desktop agents and a Security Review bot that comments on every PR with severity, attack path, and fix.
>
> New Cursor iOS app lets you monitor and message local agents running on your desktop — tap a computer in the app, approve the pairing request, and interact with any listed agent.
>
> Enables `Keep this computer awake` setting under Remote Control in Cursor's desktop settings to prevent sleep during remote agent sessions.
>
> Enterprise admins can enable remote control via `Org settings > Security and identity > Remote control`.
>
> New Security Review feature automatically reads every pull request in codebase context and posts a single review comment flagging exploitable bugs with severity, attack path, and proposed fix.
>
> Security Review supports team rules (Enterprise) to enforce codebase-specific policies.
>
> +2 more

WAS THIS USEFUL? 👍 👎

Cline

Feedback ⓘ v4.1.23 NOTES ↗ CODE ↗ RANK

Autonomous coding agent as an SDK, IDE extension, or CLI assistant.

> Cline expanded its agent toolset with new `skills`, `submit_and_exit`, `search_codebase` and `toolRoutingRules` surfaces, made MCP settings and CLI log paths configurable via environment variables, and shipped desktop-side inline Mermaid diagram rendering alongside a refreshed model catalog, a documented `--yolo` CLI flag, and a longer MCP server startup timeout.

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

01 `cline doctor log` honors `CLINE_LOG_PATH`

IMPROVED

90

`cline doctor log` now opens the file the CLI actually writes logs to by honoring the `CLINE_LOG_PATH` environment variable, instead of always defaulting to `~/.cline/data/logs/cline.log`.

Redirect CLI logs to a custom path and confirm `cline doctor log` opens that file — useful when the default data directory is on a slow or read-only volume.

```
$ export CLINE_LOG_PATH=/var/log/cline/cline.log
cline doctor log
```

— Exact command, env var, and before/after default path given. v4.1.23

02 New session tools: skills, submit_and_exit, search_codebase, toolRoutingRules

NEW INTEGRATION NEW

80

Cline adds a `skills` tool to invoke configured skills within a session, a `submit_and_exit` tool to submit a final answer and stop a session programmatically, a `search_codebase` tool to search the workspace, and a `toolRoutingRules` SDK session option to override which tools (`run_commands` or `search_codebase`) are offered per model or provider.

Override default tool routing so that `run_commands` is offered to all models on the openai-native provider, regardless of defaults.

```json
{ "toolRoutingRules": { "openai-native": ["run_commands",
"search_codebase", "fetch_web_content", "skills"] } }
```

— Names each tool and option with a config example, limited mechanism detail.

product docs

03 **Configurable MCP settings file location** NEW 70

The Cline MCP settings file, stored by default at `~/.cline/data/settings/cline_mcp_settings.json`, can now be overridden via the `CLINE_MCP_SETTINGS_PATH` environment variable.

— Names exact env var and default path but gives no usage example. product docs

04 **`-y`/`--yolo` flag documented** IMPROVED 70

The `-y` / `--yolo` flag now appears in `cline --help` output, with a warning to use it only in sandboxed environments since it skips manual tool-approval prompts.

Run Cline autonomously in a sandboxed CI environment without manual tool-approval prompts.

```
$ cline --yolo "Run tests and fix any failures"
```

— Exact flag and runnable example, minimal mechanism detail beyond that. cli-v3.0.69

05 **Inline Mermaid diagram rendering in chat** NEW 65

Cline desktop renders Mermaid diagrams inline in chat with copy, download, fullscreen, and pan/zoom controls; links inside diagrams go through the same confirmation dialog as other links.

— Describes controls and link behavior but no exact UI path. desktop-v0.0.44

06 **Model catalog refresh across providers** UPDATED INTEGRATION IMPROVED 60

Adds GPT-6.1 Sol to the recommended list and Solar Mini 4 to the free list, drops DeepSeek V4.1 Flash and space-bunny-alpha from the free list, and updates default models for 302.AI, AKI.IO, Blue Claw, CoralBricks, CrossModel, DevPass, DigitalOcean, GMI Cloud, LLM Gateway, Mistral, NanoGPT, Neon, Nvidia, Ofox, Requesty, Subconscious, The Grid AI, Vivgrid, and Vultr.

— Lists every provider and model change but gives no usage step. v4.1.23

07 **Generate Commit Message respects `.clinerules`** IMPROVED

60

The Generate Commit Message feature now applies project conventions for language, format, and ticket references defined in `.clinerules` when drafting commit messages.

— Names the config file and effect but shows no example. v4.1.23

08 **`cline config --json` output fixed** IMPROVED

60

Fixes `cline config --json` so the full Cline configuration can be inspected as machine-readable JSON, e.g. piped to `jq` to query specific fields.

Inspect the full Cline configuration as machine-readable JSON — pipe to jq to query specific fields.

```
$ cline config --json | jq '.mcp'
```

— Exact command and example but no detail on what was broken. cli-v3.0.69

THINNER COVERAGE BELOW

09 **Sessions from local/self-hosted providers without API keys**

UPDATED INTEGRATION NEW

55

Local and self-hosted providers that require no API key — LM Studio, Ollama, vLLM, LiteLLM, and custom OpenAI-compatible endpoints — can now start Cline sessions.

— Names each provider but gives no setup steps. desktop-v0.0.44

10 **MCP server startup timeout raised to 10 seconds** IMPROVED

55

The MCP server startup timeout was raised from 3 seconds to 10 seconds, so servers launched with `npx` or `uvx` on Windows load reliably instead of being silently dropped; the same 10-second timeout applies in the CLI.

— Gives old and new timeout values, no further config to adjust. cli-v3.0.69 desktop-v0.0.44

11 **Vultr model IDs renamed (breaking)** BREAKING

50

Vultr model IDs were renamed upstream as part of the model catalog refresh; any pinned Vultr model will need to be re-selected after upgrading.

— States impact and required action, no exact old/new IDs. v4.1.23

└──▷ BREAKING ON UPGRADE

! Vultr model IDs were renamed upstream; any pinned Vultr model will need to be re-selected after upgrading.

sdk/sdk/v0.0.91 NOTES ↗ CODE ↗ STABLE

Adds Langfuse BYOK tracing env vars, `AgentModelFinishReason.unknown` handling, 10s MCP stdio init budget, new rules exports, and Anthropic endpoint guard.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch sdk/sdk/v0.0.91
https://github.com/cline/cline.git
# already have the repo? check out this version:
$ git checkout sdk/sdk/v0.0.91
```

Adds opt-in Langfuse tracing for BYOK providers via `CLINE_LANGFUSE_ALL_PROVIDERS`, with operator tags, metadata, and environment via `CLINE_LANGFUSE_TAGS`, `CLINE_LANGFUSE_METADATA`, and `LANGFUSE_TRACING_ENVIRONMENT`.

Adds `AgentModelFinishReason.unknown` — the AI SDK adapter maps unified `other` and missing finish reasons to it instead of `stop`; without tool activity the agent retries once with a hidden user message before failing the run.

Raises the stdio MCP server `initialize` timeout default from 3s to 10s, allowing `npx` / `uvx` launchers through cmd.exe on Windows (which routinely take 3–6s) to complete; an explicit `timeout` still overrides it.

Adds `listEnabledRulesFromRecords` and `loadRulesForSystemPromptFromRecords` exports for rules; combined user-instruction service now orders rules by name, matching the single-workspace path.

Adds `isOfficialAnthropicEndpoint` helper to restrict Anthropic's server-side refusal `fallbacks` option to `api.anthropic.com` only, preventing 400 errors on custom endpoints such as Azure AI Foundry.

WAS THIS USEFUL? 👍 👎

yetone magpie

Feedback ⓘ v0.1.1100 CODE ↗ RANK

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

> magpie's standout addition this window is per-account Codex credit-spend control via a new `magpie quota credits` CLI command with matching API and UI toggles, alongside gateway-key migration support for clients moving from another proxy, browser-based gateway access via CORS, detection of six more agents inside WSL, and a session-list cap raised from 200 to 5000.

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

01 **Per-account Codex credit-spend control** NEW 95

Adds a `magpie quota credits [<account>] [on|off]` CLI command to toggle whether a Codex account spends its credits once its usage window is exhausted. The opt-out is stored as `codexNoCredits` in magpie's settings file, configurable via `POST /api/settings/codex-credits`, and surfaced as a Use credits toggle on each named Codex account in the Usage card and menu bar panel.

Prevent a specific Codex account from spending credits when its usage window runs out — other accounts and fallbacks absorb the traffic instead.

```
$ magpie quota credits user@example.com off
```

— Runnable CLI command, endpoint, config key, and UI toggle all named v0.1.1091

02 **Paginated sessions list in API and UI** IMPROVED 86

`GET /api/sessions` now supports `?limit=N` and returns up to 5000 sessions by default instead of being silently capped at 200. The Sessions list in the GUI, TUI, and `magpie sessions` CLI is now as long as the session count shown beside it, displayed in pages.

Retrieve more than 200 sessions from the API — useful in scripts or dashboards that need the full session history.

```
$ curl 'http://localhost:<port>/api/sessions?limit=500'
```

— Exact endpoint, param, and limit numbers with runnable example v0.1.1091

03 **User-supplied secrets for gateway keys** NEW 78

Gateway keys can now be created with a user-supplied `Secret` value via `access.Change{Secret: "..."}`, letting clients migrating from another gateway (e.g. CLIProxyAPI) keep using their existing key without re-keying. A new Named() function in `internal/access` recognizes both `sk-magpie-…` managed keys and these user-supplied keys so requests are correctly counted and rate-limited under the right key identity.

– Names exact API and package but no CLI/UI entry point shown v0.1.1100

04 **Browser access to gateway via CORS origins** NEW 78

Adds `settings.CORSOrigins` (configurable under Settings › Network) so user-listed web pages can call the gateway directly from a browser, authenticated with a gateway key, enabling browser-based integrations.

– Exact config key and UI path, lacks a worked example v0.1.1095

05 **Unofficial plugin discovery via GitHub topic** NEW 70

The Plugins page gains a Discover section that surfaces any GitHub repository tagged with the `magpie-plugin` topic as an unofficial, unreviewed plugin, installable from its npm package name or a `github:owner/repo` spec.

– Names UI location and exact install spec formats v0.1.1099

06 **Gemini image support via functionResponse** IMPROVED 65

Gemini `functionResponse` parts now reach the upstream, enabling image-capable agents (e.g. `view_file`) to see actual images via the gateway.

– Names exact field and example tool, no setup step needed v0.1.1091

07 **WSL detection for six more coding agents** NEW 60

magpie now detects and wires CodeBuddy Code, Cline, Gemini CLI, AtomCode, Goose, and Reasonix when installed inside a WSL distro, routing them through the local gateway the same as native agents.

– Names every agent covered but gives no setup steps v0.1.1097

THINNER COVERAGE BELOW

08 **New-skill discovery during update checks** NEW 50

Checking for updates now lists skills a repository has added since installation, shown beside existing library skills, with Add or Ignore actions.

– Names UI flow and actions, no exact command or key v0.1.1091

09 **Redaction bypass for local-network models** NEW 45

Allows a model running on the local computer or local network to be set to receive requests unmasked by Settings' redaction, rather than having prompts and responses redacted as with remote models.

— Describes behavior but no exact setting name or path `v0.1.1091`

10 **Model list ordering extended to every agent** IMPROVED

35

The model-list `Order` field, previously available only to Codex, is now extended to every agent.

— Thin description beyond field name and scope change `v0.1.1095`

WAS THIS USEFUL? 👍 👎

SST OpenCode

Feedback ⓘ v1.18.35 NOTES ↗ CODE ↗ RANK

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

> OpenCode expanded its Zen model marketplace with five new models and improved its stats layer with new agent-readable data formats.

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

01 **New Zen model marketplace additions** NEW

70

Adds `gpt-6.1-sol` (GPT 6.1 Sol) with tiered pricing at ≤272K and >272K token context windows, `mistral-large-4` (Mistral Large 4) at 50% off promotional pricing, and three limited-time feedback-collection models: `exo-free` (Exo Free), `fledge-alpha-free` (Fledge Alpha Free), and `ling-3.1-flash-free` (Ling 3.1 Flash Free).

— Names five models and pricing but no config key to select them `product docs`

THINNER COVERAGE BELOW

02 **Agent-readable stats formats and redirects** IMPROVED

40

Adds canonical redirects and JSON and Markdown data formats for agent-readable stats.

— States formats added but no endpoint or usage path given `v1.18.35`

WAS THIS USEFUL? 👍 👎

Graphify Labs Graphify

Feedback ⓘ v0.9.78 NOTES ↗

RANK

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

> Graphify's latest release adds versioning metadata to its graph output format, helping downstream consumers track format changes.

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

01 **Metadata fields in graph.json** NEW 60

`graph.json` now includes `schema_version` and `graphify_version` fields so downstream consumers can detect format changes and identify the producing release. The fields are written additively, so existing readers and older graphs are unaffected.

— Names exact fields and compatibility behavior, but no usage example. v0.9.78

WAS THIS USEFUL? 👍 👎

Anthropic Claude Code

Feedback ⓘ v2.1.292 NOTES ↗

CODE ↗ RANK

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

> Claude Code v2.1.292 streamlines plugin installation, adds sub-agent effort control and retry tuning, expands mod hook capabilities, and changes the default MCP protocol version for local servers.

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

01 **One-step plugin install from marketplace** NEW 80

The `--marketplace <source>` flag added to `claude plugin install` registers the marketplace (under the same policy checks as `claude plugin marketplace add`) and installs the plugin in a single command, rather than requiring a separate marketplace-add step first.

Install a plugin directly from a marketplace in one command, skipping a separate `marketplace add` step.

```
$ claude plugin install --marketplace
https://marketplace.example.com my-plugin
```

— Names flag and subcommand with a runnable install example v2.1.292

02 **Tunable overloaded-retry backoff delay** NEW 75

The new `CLAUDE_CODE_OVERLOADED_RETRY_BASE_DELAY_MS` environment variable lets users tune the backoff base delay used when retrying overloaded (529) API responses.

Give overloaded-API retries a longer base delay (e.g. 5 seconds) to reduce hammering a busy endpoint in CI.

```
$ export CLAUDE_CODE_OVERLOADED_RETRY_BASE_DELAY_MS=5000
claude -p 'run the full test suite and summarise failures'
```

— Named env var with a concrete example value v2.1.292

03 **Default MCP protocol version bump for stdio servers** BREAKING 75

Local (stdio) MCP server connections now negotiate protocol version `2026-07-28` by default; servers that do not support this version need `MCP_PROTOCOL_NEGOTIATION=legacy` set to restore prior behaviour.

Opt a local MCP server back to the legacy protocol when it does not support the new 2026-07-28 handshake.

```
$ export MCP_PROTOCOL_NEGOTIATION=legacy
claude
```

— Names exact protocol version, env var, and migration path v2.1.292

04 **New mod hooks for autocomplete, caching and workflow agents**

NEW 60

Mods gained three new hook points: a `prompt.autocomplete` mod event lets a mod inject its own autocomplete rows into the prompt box; `$.model.complete` now supports prompt caching for mods, where `prompt` and `system` accept text blocks and setting `cache: true` on a block caches the request up to that point; and the `agent.spawn` mod hook now includes workflow agents with their run and index, so a mod can inspect and refuse them.

— Names all three hooks and their fields, but no usage examples v2.1.292

THINNER COVERAGE BELOW

05 **Effort level for sub-agents** NEW 40

An `effort` parameter added to the Agent tool lets Claude launch a sub-agent at a specified effort level.

— Names the parameter but gives no usage example or values v2.1.292

└──▷ **BREAKING ON UPGRADE**

! Local (stdio) MCP server connections now negotiate protocol version `2026-07-28` by default; servers that do not support this version will need `MCP_PROTOCOL_NEGOTIATION=legacy` to restore prior behaviour.

WAS THIS USEFUL? 👍 👎

claude-mem

Feedback ⓘ v13.34.0 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 v13.34.0 adds first-party agent integrations for Pi and DeepSeek Harness, each with their own memory-recall tools and installers, plus a durable retry mechanism for native prompt entries.

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

01 Pi integration with memory recall tools NEW

93

Installable via `npx claude-mem install --ide pi`, the extension installs under `~/.pi/agent/extensions/claude-mem` and respects the `PI_CODING_AGENT_DIR` environment variable. It exposes `mem_search`, `mem_timeline`, and `mem_get_observations` tools for progressive recall inside Pi sessions. Pi now appears in agent detection and CLI selection with preselection in the interactive installer, the native installer validates required bundle, attribution, and configuration inputs before modifying the destination, installation rolls back if a file replacement fails, and repeat installation refreshes the complete extension file set. Use `npx claude-mem pi status` to check extension-file presence and capture setup.

Install the claude-mem Pi extension so memory recall is available inside Pi sessions.

```
$ npx claude-mem install --ide pi
```

Check Pi extension-file presence and capture setup after installation.

```
$ npx claude-mem pi status
```

— Names install command, path, env var, tools, status check v13.34.0

02 DeepSeek Harness integration with remote worker support NEW

93

Installable via `npx claude-mem install --ide dsh --dsh-profile web`, where `--dsh-profile` also accepts existing profiles such as `tui`. It exposes `mem_search`, `mem_timeline`, `mem_get_observations`, `mem_save`, and `mem_context` tools inside DeepSeek Harness sessions, and can target a remote claude-mem worker for recall via `DSH_MEM_BASE_URL` or plugin configuration without running a local worker. DeepSeek Harness appears in agent detection and CLI selection, its native installer validates required inputs before modifying the destination,

installation records managed profiles for later uninstall while leaving unrelated watches and host configuration in place, and it preserves `CLAUDE_MEM_TRANSCRIPTS_ENABLED=false` and the configured transcript-watch path when installed.

Install the DeepSeek Harness plugin using the default web profile, then point recall at a remote worker instead of running one locally.

```
$ npx claude-mem install --ide dsh --dsh-profile web
```

— Names install command, profiles, tools, env vars v13.34.0

THINNER COVERAGE BELOW

03 **Durable native prompt retries** IMPROVED 45

Native user-entry IDs are now stored atomically with the prompt and scoped to the worker's platform/session identity, so retrying the same entry returns the existing prompt receipt rather than creating a duplicate.

— Explains mechanism but no named flag or command v13.34.0

WAS THIS USEFUL? 👍 👎

Diagram Design

Feedback ⓘ changes since 2026-09-10 CODE ↗

Diagram Design is a Claude Code skill that generates editorial HTML and SVG diagrams and imports draw.io or Mermaid sources.

Diagram Design: Adds architecture delta diagrams.

└──▷ **TRY IT**

Run the new marimekko geometry verifier against a marimekko SVG to confirm column widths, segment heights, and area encoding all match declared amounts.

```
$ python3 scripts/verify-marimekko.py path/to/marimekko.svg
```

Adds example axonometric diagrams for AI agent stack, keyboard, coffee shop, and warehouse layouts.

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

browser-use

Feedback ⓘ 0.13.11 NOTES ↗ RANK

The browser-use Python library automates web browsers for AI agents.

> browser-use expanded its Cloud API with richer billing data and a new runs-listing endpoint, and added a toolsets_for_claude integration letting Claude drive a browser directly.

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

01 **List runs endpoint in Cloud API v4** NEW 95

Adds `GET https://api.browser-use.com/api/v4/runs` to list runs, authenticated via the `X-Browser-Use-API-Key` header, returning fields including `id`, `task`, `title`, `model`, `contextLimit`, `status`, `result`, `error`, `sessionId`, `workspaceId`, `totalInputTokens`, `totalOutputTokens`, `totalCostUsd`, `createdAt`, `updatedAt`, `modelParams`, `output`, `outputSchema`, `attachedFileIds`, and `judgement`, with cursor-based pagination via `nextCursor` and `hasMore`.

— Exact method, path, auth header and response schema given product docs

02 **Claude browser automation via toolsets_for_claude** NEW 93

Adds the `browser_use.integrations.toolsets_for_claude` module exposing `BrowserUse` and `Bash` drivers so Claude can control a local browser, a Browser Use Cloud browser, or a remote CDP browser through Anthropic's tool-runner interface; requires an Anthropic SDK release that includes `anthropic.tools.browser`.

Let Claude autonomously browse the web using Browser Use as the backend driver — wiring up both the `BrowserUse` and `Bash` toolsets so the model can navigate pages and process data in a single run.

```python
import os

from anthropic import AsyncAnthropic
from browser_use.integrations.toolsets_for_claude import Bash,
BrowserUse

task = 'Open example.com and report its page title.'
driver = BrowserUse()  # Or BrowserUse(use_cloud=True)
bash = Bash(output_dir='outputs')

async with driver, AsyncAnthropic() as client:
    runner = client.beta.messages.tool_runner(
        model=os.environ['ANTHROPIC_MODEL'],
        max_tokens=32_768,
        max_iterations=100,
        tools=[driver, bash],
        messages=[{'role': 'user', 'content': task}],
    )
    result = await runner.until_done()
```

— Runnable Python example wires both drivers end to end `0.13.11`

03

Billing API exposes credit, rate limit and plan details

IMPROVED

83

The billing endpoint (`GET https://api.browser-use.com/api/v2/billing`) response now includes `totalCreditsBalanceUsd` , `monthlyCreditsBalanceUsd` , and `additionalCreditsBalanceUsd` for credit breakdown; `rateLimit` , `concurrentSessionLimit` , and `activeSessionCount` for live throttling/session data; a `planInfo` object with `planName` , `subscriptionStatus` , `subscriptionId` , `subscriptionCurrentPeriodEnd` , and `subscriptionCanceledAt` ; and per-project/key fields `projectId` , `name` , `isFreeTier` , `apiKeyId` , and `tracingDisabled` .

Check live session usage and credit balance to gate automation jobs before launching new browser sessions.

```
$ curl -s -H 'Authorization: Bearer <your_api_key>'
https://api.browser-use.com/api/v2/billing | python3 -c "
import sys, json
d = json.load(sys.stdin)
print('Credits remaining (USD):', d['totalCreditsBalanceUsd'])
print('Active sessions:', d['activeSessionCount'], '/',
d['concurrentSessionLimit'])
print('Rate limit:', d['rateLimit'])
print('Plan:', d['planInfo']['planName'], '|', d['planInfo']
['subscriptionStatus'])
"
```

— Every new field named, example queries the live endpoint `product docs`

THINNER COVERAGE BELOW

04 **Binary image support in filesystem write_file** IMPROVED

44

Extends `write_file` in the filesystem toolset to create small binary images (PNG, GIF, JPG, WEBP) for use in upload flows.

— Names formats but no example or limits given `0.13.11`

WAS THIS USEFUL? 👍 👎

Vercel AI SDK

Feedback ⓘ ai@7.0.130 NOTES ↗ RANK

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

> AI SDK 7.0.129–7.0.130 added a portable `max` reasoning level with provider-specific mappings and fallback behavior, deduplicated deprecation warnings, and structured refusal reporting for provider decision refusals.

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

01 **Portable `max` reasoning level** BREAKING 80

Adds a portable `max` reasoning level to `LanguageModelV4CallOptions['reasoning']` and the derived `ReasoningLevel` type, with native provider mappings, compatibility coercions, and budget-based fallback support across Anthropic, OpenAI, Fireworks, and Amazon Bedrock. This widens `LanguageModelV4CallOptions['reasoning']` and `ReasoningLevel`, so third-party providers and consumers with exhaustive switches or `Record<ReasoningLevel, ...>` mappings must handle `max` before upgrading; Amazon Bedrock Nova 2 calls using portable `reasoning: 'xhigh'` now send `maxReasoningEffort: 'high'` instead of the previously unsupported `'max'` and return a compatibility warning.

– Names type, providers, and before/after mapping behavior ai@7.0.129

THINNER COVERAGE BELOW

02 **Refusal answers for provider decision refusals** NEW 40

Reports provider decision refusals as structured refusal answers via `@ai-sdk/provider@4.0.24`.

– Names the package version but mechanism is thin ai@7.0.130

03 **Deduplicated deprecation warnings with stable codes** IMPROVED 30

Emits deprecation warnings with stable codes and deduplicates them so each warning surfaces only once per process.

– Describes behavior but no code or config key named ai@7.0.129

└──▷ **BREAKING ON UPGRADE**

! Adding `max` widens `LanguageModelV4CallOptions['reasoning']` and the derived `ReasoningLevel` type — third-party providers and consumers with exhaustive switches or `Record<ReasoningLevel, ...>` mappings must handle `max` before upgrading.

! Amazon Bedrock Nova 2 calls using portable `reasoning: 'xhigh'` now send `maxReasoningEffort: 'high'` instead of the previously (unsupported) value `'max'`, and return a compatibility warning.

WAS THIS USEFUL? 👍 👎

EXXETA exxperts

Feedback ⓘ v0.14.2 NOTES ↗ RANK

The exxperts framework runs local-first AI agents with approval-gated memory, model-provider integrations, MCP tools, and web search.

> exxperts v0.14.2 hardens Memorize against sorting failures, adds conflict resolution for notes, lets Skills be imported from private git repositories, and sandboxes connectors for rooms opened on untrusted folders.

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

01 **Connector sandbox for untrusted room folders** NEW 65

Opening a room on a folder from an external source can no longer execute programs on the computer — only connectors explicitly switched on for that room in Settings > Connectors are used, blocking any connector not enabled.

Prevent an untrusted downloaded project folder from running arbitrary programs when you open it as a room.

> 📍 Open the room on the downloaded project folder, then go to Settings › Connectors for that room and switch on only the connectors you want active — any connector not explicitly enabled is blocked from running.

— Names exact settings path and security behaviour with example v0.14.2

02 **Note conflict resolution and Keep vs pin protection** NEW 60

When a new note lands beside a pinned note or may conflict with an older one, exxperts now prompts Keep both or Replace. Notes record when they were learned so an older conversation can never overwrite a newer note — the superseded value is preserved in the Archive as history. Keep now only protects a note for a single Memorize or Review update; permanent protection requires pinning the note.

— Clear mechanism and named states, no direct command v0.14.2

03 **Long conversation handling in rooms** IMPROVED 60

Rooms now open and save very long conversations, showing their newest 2,000 messages. When a room reaches its remembered-conversation limit, its Home card displays a notice and Maintain opens so you can run Memorize.

— Names concrete limit and the Home/Maintain navigation path v0.14.2

THINNER COVERAGE BELOW

04 **'Memorize without the card' renamed and expanded** BREAKING

55

The Memorize without the card setting is renamed 'Memorize without the first read or the card' and now also skips the first read — Start runs immediately and Discuss first opens the first read. Workflows relying on the first read running under the old setting will no longer get it.

— Names exact old and new settings and behaviour change v0.14.2

---

05 **Skills import from private git repositories** NEW 55

Skills can now be imported from a private repository using the git credentials already stored on the computer — no separate authentication step is required.

Import a Skill from a private git repository using credentials already stored on the machine — no extra authentication step required.

> 📍 In the Skills panel, choose Import, paste the private repository URL, and confirm — exxperts uses the git sign-in already on this computer to fetch it.

— Has a usage example with exact steps v0.14.2

---

06 **Fault-tolerant Memorize sorting** IMPROVED 50

When Memorize hits a sorting failure, it now keeps the conversation summary in Unsorted instead of aborting, and continues — a later Memorize or Review run can then sort it into topics.

— Explains recovery mechanism but no command to trigger it v0.14.2

└──▷ **BREAKING ON UPGRADE**

! The Memorize without the card setting is renamed to 'Memorize without the first read or the card' and now also skips the first read — any workflow that relied on the first read running under that setting will no longer do so.

WAS THIS USEFUL? 👍 👎

OpenWorker

Feedback ⓘ v0.3.4 CODE ↗ RANK

OpenWorker automates desktop tasks across files, tools, and integrations with governed, audited AI agent actions.

> OpenWorker split OpenRouter into separate API-key and browser-authenticated account providers (with a breaking route change), added llama.cpp and vLLM as first-class local providers, per-model configuration, machine-aware Ollama context sizing, and a round of Composer and Models & Keys UI upgrades.

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

01 OpenRouter API-key and account authentication

BREAKING

90

OpenRouter integration offers two separate provider cards under Settings → Models & Keys : `OpenRouter` (API keys, paste a key, with a `Manage or revoke keys on OpenRouter` link) and `OpenRouter account` (Subscriptions, browser sign-in via an S256 PKCE flow using your OpenRouter credits). On a remote backend, Use a manual code lets you paste the code OpenRouter displays to authenticate without exposing your key. In v0.3.4 the `openrouter` provider reverted to a plain API-key provider — the `auth_method` and `account_connected` flags and related credential logic were removed — and sign-in routes moved from the old `openrouter` card paths to `/v1/providers/openrouter-account/…`, breaking any OAuth callback integration that relied on the old paths.

Authenticate to OpenRouter from a remote machine where browser sign-in is unavailable.

> 📍 In OpenWorker go to Settings → Models & Keys, select the OpenRouter account card, choose Use a manual code, then paste the code shown on OpenRouter's website.

Sign in to OpenRouter with your account (using PKCE browser flow) so OpenRouter credits are used automatically, without pasting an API key.

> 📍 Settings → Models & Keys → Subscriptions → OpenRouter account → Sign in with OpenRouter → authorise in browser → return to app.

— Covers full auth mechanism plus the breaking route and flag removal. v0.3.4

02 Remote backend server launch command

NEW

90

Launches the OpenWorker backend server remotely with `.venv/bin/python -m coworker.server.run --cwd <project> --port <port>` (Windows: `.venv\Scripts\python -m coworker.server.run`), letting a team member connect without sharing a raw API key.

Run the OpenWorker backend remotely so a team member can connect without sharing a raw API key.

```
$ .venv/bin/python -m coworker.server.run --cwd ~/some/project --port 8765
```

— Exact runnable command with flags given. product docs

03 **Per-model settings file** NEW 85

Adds per-model settings stored in `model_config.json` in the state folder, controlling context size, longest reply, thinking on/off, sampling, auto-compact, and which model new sessions start with; unset values are filled in from a recommended-settings table in `coworker/providers/recommended_models.json`.

— Named config files but no edit workflow shown. v0.3.4

04 **llama.cpp and vLLM local providers** NEW 75

Adds llama.cpp (`localhost:8080` by default, `/v1` compatible) and vLLM (`localhost:8000` by default) as first-class local providers, reachable from Settings → Models & Keys, alongside existing providers.

Connect a locally running llama.cpp server as a provider and pick its model for a session, without touching the OpenAI custom-endpoint field.

📍 Settings → Models & Keys → llama.cpp → enter server address (default: localhost:8080) → pick the loaded model → open Composer and select it from the grouped model picker.

— Default ports named with a connect-and-select walkthrough. v0.3.4

05 **Models & Keys settings page overhaul** NEW 70

Adds a grouped provider layout and local-model table to the Models & Keys settings page, including a system-info panel showing what the machine is and what Ollama's models can do, a live connectivity indicator for keyless local providers (llama.cpp, vLLM, Ollama) showing alive status based on cached server probes, and a Retry button on the local model list when the local server cannot be reached.

— Names several UI additions and the probing mechanism, no exact steps. v0.3.4

06 **Machine-aware Ollama context sizing** IMPROVED 65

Adds machine-aware Ollama context sizing (16K / 32K / 64K / up to 128K) based on available GPU or system memory, replacing Ollama's default 4K tier.

— Automatic behavior change with no user-facing action. v0.3.4

07 **Composer model and session controls** NEW 65

Adds per-session model controls in Composer: a reasoning-effort and thinking controls panel beside the model picker, a grouped model picker for selecting among providers and local models in a single session, and typed-answer support in the main message box for open `ask_user` questions on Enter, with placeholder text 'Type your answer — or pick an option above'.

— UI locations described but no exact navigation steps. v0.3.4

└──▷ **BREAKING ON UPGRADE**

! The `openrouter` provider reverts to a plain API-key provider; `auth_method` and `account_connected` flags and all special-case OpenRouter credential logic are removed — any configuration or code that relied on those flags will break.

! OpenRouter sign-in routes move from the previous `openrouter` card paths to `/v1/providers/openrouter-account/…`; existing OAuth callback integrations must be updated to the new path.

WAS THIS USEFUL? 👍 👎

CopilotKit OpenBot

Feedback ⓘ v0.1.1 NOTES ↗ RANK

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

> OpenBot v0.1.1 tightens network and database security — adding TLS sslmode support for managed Postgres, closing loopback-address and localhost-subdomain bypasses, and gating coordination tools in group conversations by actual permission.

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

01 **TLS sslmode options on DATABASE_URL** BREAKING 90

`DATABASE_URL` now supports `sslmode=require`, `sslmode=disable`, and `sslmode=verify-full` (with optional `sslrootcert`) so deployments on RDS, Cloud SQL, or Azure Database can establish encrypted Postgres connections. `verify-ca`, `prefer`, and `allow` are now refused with an error naming the supported modes, and `sslmode=verify-full` requires Bun 1.4 or later, being refused on Bun 1.3.

Connect an OpenBot deployment on RDS or Cloud SQL to its managed Postgres over TLS, verifying the server certificate.

```text
DATABASE_URL=postgres://user:pass@db.example.com:5432/mydb?
sslmode=verify-full&sslrootcert=/etc/ssl/certs/rds-ca.pem
```

Require an encrypted but unverified connection to a managed Postgres instance when running on Bun 1.3.

```text
DATABASE_URL=postgres://user:pass@db.example.com:5432/mydb?
sslmode=require
```

— Names exact flag, values, and Bun version requirement with examples v0.1.1

02 **Stricter loopback-address blocking for single-user mode**

70

BREAKING

`OPENBOT_SINGLE_USER` now refuses public addresses that merely begin with `127.` (such as `https://127.0.0.1.nip.io`) — only a full `127.x.x.x` address now counts as loopback, so deployments relying on the looser match will fail to start.

— Names exact env var and example address, no migration guidance `v0.1.1`

THINNER COVERAGE BELOW

03 **Localhost subdomain blocking in Bot browser** IMPROVED

55

A Bot's browser now refuses names under `.localhost` (e.g. `http://admin.localhost:5432`), matching the existing block on bare `localhost`.

— Describes exact pattern blocked but lacks mechanism detail `v0.1.1`

04 **Permission-gated coordination tools in group conversations**

NEW

50

`ask_person` and `message_bot` coordination tools are now offered to a Bot's turn in a group conversation only when a call from that run would actually be permitted.

— Names the two tools but gives no usage example `v0.1.1`

└──▷ **BREAKING ON UPGRADE**

! `sslmode=verify-ca`, `prefer`, and `allow` values on `DATABASE_URL` are now refused with an error naming the supported modes — deployments using those values will fail to start.

! `sslmode=verify-full` on `DATABASE_URL` requires Bun 1.4 or later and is refused on Bun 1.3.

! `OPENBOT_SINGLE_USER` deployments whose public URL begins with `127.` (e.g. `https://127.0.0.1.nip.io`) are now refused — those deployments will fail to start.

WAS THIS USEFUL? 👍 👎

Brig

Feedback ⓘ v0.4.0 NOTES ↗ RANK

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

> Brig v0.4.0 adds `brig plan` for dry-run previews and `brig sh -c` for inline scripts, extends egress enforcement to nerdctl, and adds boot-asset digest verification, alongside a cluster of breaking changes covering runtime shim selection, exit codes, sandbox networking defaults, and profile file permissions.

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

01 **BRIG_CONTAINERD_RUNTIME shim restriction** BREAKING 95

`BRIG_CONTAINERD_RUNTIME` no longer accepts a shim that shares the host kernel. Values that name `runc`, `crun`, or a shim containerd resolves to `io.containerd.runc.v2` are refused at boot or join. Unset the variable to use the default `io.containerd.urunc.v2` shim. Stop any sandbox still running under a refused shim with `brig stop <ref>` before the next `brig run`.

— Names exact env var, shim values, and migration command v0.4.0

02 **Profile files: binding mode quoting required** BREAKING 80

A profile `files:` binding with an unquoted mode (e.g. `644`) is now refused; quote the mode (e.g. `"0644"`). A file that overrides a built-in profile using an unquoted mode is ignored with a warning and the built-in profile runs instead.

— Shows exact before/after syntax and fallback behavior v0.4.0

03 **brig sh argument handling and inline scripts** BREAKING 75

Adds `brig sh -c` to run an inline script inside a sandbox guest. `brig sh <ref> <words...>` now passes each word as a literal argument — shell parsing and variable assignments such as `FOO=bar` in front of the command no longer work. Use `brig sh <ref> -c '<script>'` for scripts and `brig sh <ref> env FOO=bar <cmd>` to set variables.

— Gives exact replacement commands for the changed behavior v0.4.0

04 **Exit code 7 for policy-enforcement failures** BREAKING 65

Exits with code `7` (previously `1`) when a runtime cannot enforce a planned property, giving scripts a distinct signal for policy-enforcement failures; scripts checking `exit 1` for this condition must be updated.

— Names old and new exit codes and required script update v0.4.0

05 **Telemetry events and controls** NEW 65

Emits brig's own telemetry events, visible and controllable via `brig telemetry status` and `brig telemetry off`.

— Exact subcommands given for checking and disabling telemetry v0.4.0

THINNER COVERAGE BELOW

06 **`brig plan` execution envelope preview** NEW 55

Adds `brig plan` to preview a run's execution envelope without decrypting secrets.

— Names the command and its secrets-safe behavior v0.4.0

07 **Boot-asset digest verification** NEW 50

Verifies that a boot asset's digest matches the bundle before booting, refusing mismatches, and checks the runtime bundle's own record as part of the verification step.

— Describes mechanism but no exact flag or command v0.4.0

08 **Sandbox network isolation by default** BREAKING 45

Sandbox networks are now isolated by default on new `brig wrap` sandboxes; sandboxes that relied on a shared network will behave differently after upgrade.

— Names the affected command but no migration flag v0.4.0

09 **Egress enforcement extended to nerdctl** IMPROVED 40

Enforces egress policies on nerdctl, expanding policy enforcement beyond the hvi backend.

— Names backends but no config surface or example v0.4.0

10 **Formatted run status notices** IMPROVED 20

Displays first-run notices, verification alerts, and run warnings as formatted short blocks.

— Generic UX description with no named surface v0.4.0

└──▷ BREAKING ON UPGRADE

! `brig wrap` now isolates new sandbox networks by default; sandboxes that relied on a shared network will behave differently after upgrade.

! `BRIG_CONTAINERD_RUNTIME` no longer accepts a shim that shares the host kernel. Values that name `runc`, `crun`, or a shim containerd resolves to `io.containerd.runc.v2` are

refused at boot or join. Unset the variable to use the default `io.containerd.urunc.v2` shim. Stop any sandbox still running under a refused shim with `brig stop <ref>` before the next `brig run`.

! Exit code for a runtime that cannot enforce a planned property changed from `1` to `7`; scripts checking `exit 1` for this condition must be updated.

! `brig sh <ref> <words...>` now passes each word as a literal argument — shell parsing and variable assignments such as `FOO=bar` in front of the command no longer work. Use `brig sh <ref> -c '<script>'` for scripts and `brig sh <ref> env FOO=bar <cmd>` to set variables.

! A profile `files:` binding with an unquoted mode (e.g. `644`) is now refused; quote the mode (e.g. `"0644"`). A file that overrides a built-in profile using an unquoted mode is ignored with a warning and the built-in profile runs instead.

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Agentic Infrastructure

Nebius AI Cloud

Feedback ⓘ changelog · Oct 7, 2026

CHANGELOG ↗ RANK

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

> Nebius AI Cloud shipped a Wiz integration for misconfiguration scanning, a new read-only security-auditor role, keyless tunnel agent authentication via service accounts, and a Kubernetes storage class parameter for large disk block sizes, alongside two breaking changes to the CLI installer URL and metadata service header handling.

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

01 Disk block size control for Kubernetes CSI volumes NEW

88

Adds the `disk_block_size_bytes` storage class parameter for CSI-provisioned disks in Managed Service for Kubernetes, enabling larger block sizes for disks over 16 TiB to avoid performance degradation at scale.

Provision a large (>16 TiB) Kubernetes persistent volume with an optimised block size to avoid performance degradation at scale.

yaml

```yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: large-disk-optimised
provisioner: disk-csi-driver.mks.nebius.com
parameters:
  disk_block_size_bytes: '65536'
```

— Named config key with a runnable StorageClass example changelog-20261007-27bc3438

02 Ubuntu 24.04 CUDA 13 boot image family for GPU VMs NEW

73

Adds the `ubuntu24.04-cuda13-latest` boot disk image family for GPU VMs, kept current with the latest GPU driver and CUDA 13.x versions; use `ubuntu24.04-cuda13.0` for a fixed version.

— Exact image names given, directly usable when launching a VM

changelog-20261007-27bc3438

03 **Keyless tunnel agent authentication via service accounts** NEW

65

Adds the `applicationtunnel.agent` permission under the Nebius Tunnels role for service accounts, enabling tunnel agents to authenticate using a VM's attached service account without creating or copying service account keys.

— Named role and permission but no step-by-step setup changelog-20261007-27bc3438

04 **Metadata service blocks spoofable forwarding headers** BREAKING

65

The instance metadata service now returns `403 Forbidden` for requests containing X-Forwarded-For, X-Forwarded-Host, X-Real-IP, or Forwarded headers; workloads or proxies that send these headers to the metadata endpoint will fail.

— Exact headers and response code named, with migration impact noted

changelog-20261007-27bc3438

05 **security-auditor role for read-only compliance access** NEW

60

Adds a `security-auditor` security role granting read-only access to resource configurations across all services without exposing data, designed for third-party security and compliance platforms.

— Names the role and scope but no assignment steps changelog-20261007-27bc3438

THINNER COVERAGE BELOW

06 **CLI installer script URL moved to artifacts.nebius.cloud**

55

BREAKING

The Nebius AI Cloud CLI installation script is now downloaded from `https://artifacts.nebius.cloud/cli/install.sh`; any automation or documentation pointing to the old URL will stop working.

— Exact new URL given, directly actionable for automation updates

changelog-20261007-27bc3438

07 **Free egress under 3x average storage on Intelligent class**

IMPROVED

50

Egress traffic from the Intelligent storage class in Object Storage is free while it remains under three times the monthly average used storage capacity.

— Concrete threshold given but no usage steps changelog-20261007-27bc3438

08 **Wiz integration for resource misconfiguration scanning** NEW

48

Adds a Wiz integration to scan Compute, Managed Service for Kubernetes, Virtual Networks, and Object Storage resources for misconfigurations.

— Names covered services but no setup or endpoint detail `changelog-20261007-27bc3438`

---

09 Observability Agent reports MSK add-on metrics IMPROVED

46

Nebius Observability Agent v1.0.9 now sends metrics from Managed Service for Kubernetes add-ons (such as Cilium and Hubble) to Observability Metrics.

— Names the version and add-ons but no config instructions `changelog-20261007-27bc3438`

└──▷ BREAKING ON UPGRADE

! The Nebius AI Cloud CLI installation script URL has changed to `https://artifacts.nebius.cloud/cli/install.sh`; any automation or documentation pointing to the old URL will stop working.

! The instance metadata service now returns `403 Forbidden` for requests containing `X-Forwarded-For`, `X-Forwarded-Host`, `X-Real-IP`, or `Forwarded` headers; workloads or proxies that send these headers to the metadata endpoint will fail.

WAS THIS USEFUL? 👍 👎

Snowflake Cortex AI

Feedback ⓘ changelog · Oct 7, 2026

CHANGELOG ↗

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

Cortex AI adds response caching for AI functions, a code execution tool for agents, Semantic Studio GA, Spark Monitoring UI, and new Account Usage and lineage surfaces.

Adds `ACCOUNT_USAGE.STREAMS` view for querying stream metadata via Account Usage.

Adds response caching for Cortex AI Functions to reduce latency and cost on repeated queries.

NEW SUPPORTED PLATFORM Adds Cortex Agent code execution tool (generally available), allowing Cortex Agents to run code as part of agentic workflows.

NEW SUPPORTED PLATFORM Adds observability for Cortex Agents within Snowflake Native Apps.

Adds `DEFINE SEMANTIC VIEW` support in DCM Projects (generally available).

+10 more

NEW INTEGRATION Anthropic Claude Sonnet 5.5 model support

NEW INTEGRATION Agentic Marketplace Discovery

NEW SUPPORTED PLATFORM Migration of gen 1 Openflow deployments and runtimes to gen 2 general availability

NEW COMPLIANCE FRAMEWORK User-managed Restricted Session Scope in CoCo general availability

WAS THIS USEFUL? 👍 👎

Thinking Machines Lab Tinker

Feedback ⓘ

changelog · Oct 7, 2026 CHANGELOG ↗ RANK

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

> Tinker SDK 0.32.0 adds a checkpoint storage billing API and CLI command, a pickle-free sampler handle for cross-process sampling, and alternative-token sampling at every prompt position, alongside new checkpoint/session UI filters and spend breakdowns, plus a breaking change to client constructor keyword arguments.

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

01 **Checkpoint storage usage API and CLI command** NEW 95

Adds RestClient.get_current_checkpoint_storage_usage() and get_current_checkpoint_storage_usage_async() to retrieve live checkpoint storage usage grouped by project and session owner, including checkpoint count, size, and `estimated_monthly_cost_usd` (a gross rate-card projection). Also adds the `tinker billing checkpoint-storage` CLI command with an optional `--project-id` flag to scope usage to a single project.

Check your org's live checkpoint storage cost breakdown, scoped to one project, before an invoice arrives.

```
$ tinker billing checkpoint-storage --project-id <your-project-id>
```

— Names both API methods, CLI command, flag, field, with runnable example.

changelog-20261007-c2cc36d6

02 **prompt_alt_tokens_k alternative token sampling** NEW 95

Adds the `prompt_alt_tokens_k` request parameter to draw alternative tokens from the model's next-token distribution at every prompt position, returned in `SampleResponse.prompt_alt_tokens` as `tokens` (int64) and `logprobs` (float32) tensors of shape `[len(prompt) - 1, k]`.

Retrieve alternative next-token candidates at every prompt position to analyse model uncertainty across the full prompt.

```python
response = sampler.sample(prompt=tokens, prompt_alt_tokens_k=5)
alt_tokens  = response.prompt_alt_tokens.tokens   # shape
[len(prompt)-1, 5], int64
alt_logprobs = response.prompt_alt_tokens.logprobs  # shape
[len(prompt)-1, 5], float32
```

— Exact parameter, return shape, and runnable example given. changelog-20261007-c2cc36d6

---

03 **Pickle-free sampler handle for cross-process sampling** NEW

90

Adds SamplingClient.from_sampler_handle(handle) and .get_sampler_handle() as a pickle-free mechanism to pass a sampling client across processes using a plain-string handle and per-process credential resolution via `api_key=`.

Hand a sampling client to a worker process without pickling credentials — each worker authenticates with its own API key.

```python
handle = client.get_sampler_handle()
worker_client = SamplingClient.from_sampler_handle(handle,
api_key="<worker-api-key>")
```

— Named methods plus runnable cross-process example. changelog-20261007-c2cc36d6

---

04 **Client constructor keyword arguments restricted and renamed**

80

BREAKING

Adds `base_url`, `api_key`, and `default_headers_overrides` as explicit keyword arguments to the client constructor, where `default_headers_overrides` replaces `default_headers`. Any other keyword argument previously forwarded to the underlying HTTP client is now silently ignored with a `FutureWarning` and will become an error in a future release, so code passing `default_headers=` must be updated.

— Names exact kwargs and migration path, but no code sample. changelog-20261007-c2cc36d6

---

05 **Checkpoint list UI filters and base model display** IMPROVED

65

Checkpoint list UI gains per-checkpoint base model display and a header filter panel supporting project, My checkpoints / All checkpoints toggle, checkpoint type, one or more base models, and key/value pairs; active filters are preserved in the URL.

— Describes UI filters in detail, no direct runnable action. changelog-20261007-c2cc36d6

06 **loss_fn_inputs extra-key support for custom loss functions**

IMPROVED 60

Adds `loss_fn_inputs` extra-key support so custom keys beyond `target_tokens` and `weights` can be passed through to user-defined loss functions without raising a `ValueError`.

— Named config keys but no usage example. changelog-20261007-c2cc36d6

07 **Session usage and spend breakdown in UI** NEW 60

Session list and detail page now show estimated usage — tokens, tokens excluding cache hits, estimated spend, and top model — plus a Spend by chart breaking cost down by usage type and base model.

— Clear UI description but no navigation path given. changelog-20261007-c2cc36d6

THINNER COVERAGE BELOW

08 **Immediate authentication error on failed credentials** IMPROVED 50

Raises `AuthenticationError` (and other credential errors) immediately on failed credential acquisition instead of retrying for minutes and reporting a connection error.

— Named error class and before/after behaviour, no example. changelog-20261007-c2cc36d6

09 **Forward-compatible billing usage event parsing** IMPROVED 30

Billing usage events now ignore unrecognised fields so future API additions do not break parsing in SDK 0.31.0.

— Thin description, no named fields or example. changelog-20261007-c2cc36d6

└──▷ **BREAKING ON UPGRADE**

! Any keyword argument to the client constructor other than `base_url`, `api_key`, and `default_headers_overrides` — previously forwarded to the underlying HTTP client — is now silently ignored with a `FutureWarning` and will become an error in a future release.

! `default_headers` constructor argument is replaced by `default_headers_overrides`; code passing `default_headers=` must be updated.

WAS THIS USEFUL? 👍 👎

Eigen Labs Darkbloom

Feedback ⓘ v0.9.18 NOTES ↗

CODE ↗ RANK

Private Inference Network on Idle Macs

> Darkbloom v0.9.18 introduces a GDPR account erasure pipeline and makes App Attest mandatory for base rewards, alongside a new consumer referral program, Autopilot model inventory reporting, and smaller additions to MTP sampling, data retention, email campaigns, and devnet tooling.

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

01 **GDPR account erasure pipeline** NEW 65

Adds a GDPR account erasure pipeline covering `plan`, soft-delete, grace period, scrub, and outbox delivery to Stripe and Datadog.

– Names pipeline stages and delivery targets but no user-facing action v0.9.18

THINNER COVERAGE BELOW

02 **App Attest required for base rewards** BREAKING 55

App Attest is now required for base rewards; providers relying solely on legacy MDM authorization will no longer qualify, and legacy MDM eligibility is frozen for new enrollments.

– Clear before/after and migration implication for providers v0.9.18

03 **Per-model typical MTP draft acceptance** NEW 45

Adds per-model typical MTP draft acceptance for sampled requests, defaulting eligible sampled MTP requests to typical acceptance.

– Describes a default behavior change with limited mechanism detail v0.9.18

04 **Devnet installer binding** NEW 40

Adds a devnet installer binding that ties dev-installed providers to their coordinator and seeds data.

– Names the mechanism but lacks usage detail v0.9.18

05 **Open Sales Program referral rewards** NEW 35

Adds an Open Sales Program with 5% consumer referral rewards.

– Names a concrete reward percentage but no mechanism detail v0.9.18

06 **Autopilot model inventory reporting** NEW 30

Coordinator now reports connected Autopilot model inventory.

— Bare description of a reporting capability, no scope given v0.9.18

07 **14-day operational retention and async history queries** NEW

25

Prepares 14-day operational retention and asynchronous history queries.

— Thin, preparatory description with a stated retention window v0.9.18

08 **Resend provider update campaigns** NEW

25

Adds Resend provider update campaigns.

— Bare feature name with no mechanism or scope v0.9.18

└──▷ **BREAKING ON UPGRADE**

! App Attest is now required for base rewards; providers relying solely on legacy MDM authorization will no longer qualify.

WAS THIS USEFUL? 👍 👎

NVIDIA Dynamo

Feedback ⓘ v1.5.1 NOTES ↗ RANK

Dynamo is a datacenter-scale distributed inference serving framework with disaggregated prefill and decode, KV-cache-aware routing, and multi-engine support.

> Dynamo v1.5.1 tightens multimodal media handling with explicit proxy opt-in and local-path restrictions, adds configurable size caps for inline and remote media, and consolidates on aiohttp as the sole HTTP client backend, alongside protocol and renderer upgrades for Claude Code support and leaner chat-completion responses.

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

01 Explicit opt-in required for proxy-routed media fetches

95

BREAKING

Adds `DYN_MM_TRUST_EGRESS_PROXY=1` so workers must explicitly opt in to fetching request media through an ambient HTTP proxy; without it, proxy-routed media fetches are refused, though fetches to hosts in `NO_PROXY` and fetches under `DYN_MM_ALLOW_INTERNAL=1` are unaffected.

Allow a worker running behind a corporate HTTP proxy to fetch remote media, while keeping the address-validation policy active.

```
$ export DYN_MM_TRUST_EGRESS_PROXY=1
```

— Flag, before/after behavior, and exceptions all named with a command. v1.5.1

02 Multimodal input size caps NEW

85

Adds `DYN_MM_MAX_DATA_URL_MB` (default 16 MiB) to cap inline `data:` media URLs across the SGLang, TensorRT-LLM, vLLM, and vLLM-Omni paths, with requests exceeding the limit returning HTTP 400; adds `DYN_MM_MAX_FILE_SIZE_MB` (default 64 MiB) to cap remote media downloads for SGLang image-diffusion and video-generation workers.

Raise the inline media URL size limit on both the frontend and workers when handling large multimodal payloads that exceed the 16 MiB default.

```
$ export DYN_MM_MAX_DATA_URL_MB=64
```

— Names both env vars, defaults, scope and the 400 response. v1.5.1

03 Allowed-directory restriction for local media references

85

BREAKING

Adds `DYN_MM_LOCAL_PATH` to restrict SGLang diffusion workers to a named allowed directory when accepting local `input_reference` paths; previously any local path was accepted, so upgraders must now set this to continue using local image/video inputs.

Restrict SGLang diffusion workers to only read local input references from a specific directory, preventing arbitrary path access.

```
$ export DYN_MM_LOCAL_PATH=/data/diffusion-inputs
```

— Names the env var, scope, and before/after with a runnable example. v1.5.1

04 **Removal of httpx HTTP client backend** BREAKING

65

Removes the opt-in httpx backend from Dynamo's HTTP client, leaving aiohttp as the sole backend; setting `DYN_HTTP_BACKEND=httpx` now logs a warning and falls back to aiohttp instead of using httpx.

— Names the config var and fallback behavior but no example. v1.5.1

05 **Optional field omission in chat completions responses**

IMPROVED

60

Upgrades `dynamo-protocols` to v5.4.1 so `/v1/chat/completions` responses omit absent optional fields such as `usage` instead of returning `null`.

— Names endpoint and exact field behavior change. v1.5.1

THINNER COVERAGE BELOW

06 **Anthropic Messages API support via Claude Code** NEW

55

Upgrades `dynamo-renderer` to v5.1.2 and `dynamo-tokenizers` to v1.8.1, adding support for Anthropic Messages API requests from Claude Code with non-leading system messages.

— Names versions and the new request format supported. v1.5.1

07 **Default reasoning effort behavior in dynamo-renderer** IMPROVED

55

Sets `dynamo-renderer` default reasoning effort to `high` when omitted or invalid, treats `none` as thinking disabled, and allows a top-level `reasoning_effort` field to override the template argument.

— Explains default and override mechanism but no usage example. v1.5.1

08 ModelExpress dependency upgrade IMPROVED 30

Upgrades ModelExpress to v0.6.0 in the SGLang and vLLM runtime images.

— Bare version bump with no further detail given. v1.5.1

└──▷ BREAKING ON UPGRADE

! Workers that fetch request media through an ambient HTTP proxy must now set `DYN_MM_TRUST_EGRESS_PROXY=1`; without it, proxy-routed media fetches are refused. Fetches to hosts in `NO_PROXY` and fetches under `DYN_MM_ALLOW_INTERNAL=1` are unaffected.

! Local image and video files passed as `input_reference` to SGLang image-diffusion and video-generation workers now require `DYN_MM_LOCAL_PATH` to name the allowed directory; previously any local path was accepted.

WAS THIS USEFUL?  

HeyGen HyperFrames

Feedback ⓘ v0.8.139 NOTES ↗ RANK

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

> HyperFrames shipped real-time word streaming for transcription, a new gooey split block for the Registry catalog, and a CLI command to catch up on desktop-app changes after project handover.

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

01 **Real-time word streaming in transcribe --json** IMPROVED

67

The `transcribe --json` progress stream now delivers heard words as they are recognized in real time; this is not supported with `parakeet-mlx`.

— Names exact flag and a stated limitation v0.8.139

THINNER COVERAGE BELOW

02 **Gooey split block in Registry catalog** NEW

45

A new gooey split block with metaball goo and directional motion blur has been added to the Registry catalog.

— Describes effect but no usage path given v0.8.139

03 **CLI command for desktop-app handover changes** NEW

35

A new CLI command lets users catch up on changes the desktop app made to a project since it was handed over.

— No command name or syntax given v0.8.138

WAS THIS USEFUL? 👍 👎

AWS Amazon Bedrock Feedback ⓘ

Added India geographic cross-Region inference for Kimi K3 (2026-10-06) BLOG ↗

Amazon Bedrock provides managed foundation-model inference, customization, knowledge bases, and APIs for generative AI applications.

NEW SUPPORTED PLATFORM Kimi K3 India geographic cross-Region inference

WAS THIS USEFUL?  

OpenAI TypeScript SDK

Feedback ⓘ v7.29.0 NOTES ↗

RANK

The OpenAI TypeScript SDK provides typed JavaScript and TypeScript clients for OpenAI APIs and streaming responses.

> OpenAI TypeScript SDK v7.29.0 adds deferred loading for tool search and expands local tool handling in the agents framework, including failure observation and session-creation execution.

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

01 **Deferred loading for tool search** NEW 65

Typed function helpers in the `tools` module now support `defer_loading=true` for lazy/deferred loading of tool search definitions.

— Names exact flag and module but no further mechanism detail v7.29.0

THINNER COVERAGE BELOW

02 **Local tool failure observation and session-start execution in agents framework**

NEW 45

The agents framework can now surface errors from local tool calls to agent event streams, and local tools can run during session creation streams, enabling tool execution at session startup.

— Describes two related agent capabilities but no API names given v7.29.0

WAS THIS USEFUL? 👍 👎

OpenAI Python SDK

Feedback ⓘ v3.25.0 NOTES ↗ RANK

The OpenAI Python SDK provides typed Python clients for OpenAI model APIs, streaming, and asynchronous requests.

> OpenAI Python SDK v3.25.0 expands agent tooling with public access to response-format conversion and several local-tool improvements in the agents subsystem, plus richer response model fields.

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

01 Public `type_to_response_format_param` export NEW

75

`type_to_response_format_param` is now exported publicly from `openai.lib`, letting callers convert Python types (e.g. Pydantic models or dataclasses) to structured response-format parameters without reaching into private internals.

Convert a Pydantic or dataclass type to a structured response-format parameter without accessing private internals.

python

```python
from openai.lib import type_to_response_format_param
from pydantic import BaseModel

class MyOutput(BaseModel):
    answer: str
    confidence: float

response_format = type_to_response_format_param(MyOutput)
print(response_format)
```

— Named export with runnable example, but no mechanism detail v3.25.0

THINNER COVERAGE BELOW

02 Local tool handling improvements in agents subsystem IMPROVED

50

The agents subsystem gains three local-tool additions: typed function tool helpers now support `defer_loading=true`, enabling tool search during agent runs without eagerly loading all tool definitions; local tools can now run during streamed session creation; and local tool failures now surface observability, exposing errors from local tool invocations.

— Names the flag and behaviors but no usage example v3.25.0

03 **Agent turn items and usage source grouping in responses** NEW

20

The API response model adds agent turn items and usage source grouping.

— Bare mention with no fields or mechanism named `v3.25.0`

WAS THIS USEFUL? 👍 👎

Unsloth AI Unsloth

Feedback ⓘ v0.1.903-beta NOTES ↗

PRE-RELEASE

Unsloth runs and fine-tunes language and diffusion models locally through a Python library and interface.

Adds a built-in browser panel, `offload_layers` for lower-VRAM LoRA training, `unsloth start --app` for desktop agent routing, and multi-model loading via `Settings > Resources`.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.1.903-beta
https://github.com/unslothai/unsloth.git
# already have the repo? check out this version:
$ git checkout v0.1.903-beta
```

Adds built-in browser panel beside chat, opening attached files (PDF, Office, HTML, Markdown, code, text), live web pages and model-written HTML as tabs.

Adds `Settings > Browser` to pick a search engine, set default zoom, and import bookmarks from Chrome, Safari, Firefox or Edge.

Adds `offload_layers` training option (opt-in, LoRA, NVIDIA and AMD) to keep frozen layers in system RAM, capping Llama-3.1-8B 4-bit training at 7 GiB VRAM.

Adds `unsloth start --app` to let the Codex, OpenCode, OpenClaw and Hermes desktop apps use your local model.

Adds `Settings > Resources` opt-in to keep multiple models loaded simultaneously, routing each request to the model it asks for.

+9 more

NEW INTEGRATION vLLM and SGLang inference engine support

NEW SUPPORTED PLATFORM OS sandbox support for Colab and Docker

NEW SUPPORTED PLATFORM AMD Ryzen AI NPU export for fine-tuned models

NEW INTEGRATION Audio models: 70+ GGUF models across Speak, Clone, Music, Transcribe, Edit, Separate, Convert

WAS THIS USEFUL? 👍 👎

Hugging Face Transformers

Feedback ⓘ v5.19.0

NOTES ↗

Transformers: the model-definition framework for state-of-the-art machine learning models in text, vision, audio, and multimodal models, for both inference and training.

Adds expert-parallel token dispatch via `ep_dispatch_experts`, per-layer cache config for `DynamicCache` / `StaticCache`, `Trainer.loss_is_scaled_for_ga`, and new processing backend registration.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v5.19.0
https://github.com/huggingface/transformers.git
# already have the repo? check out this version:
$ git checkout v5.19.0
```

Adds `ep_dispatch_experts` plan rule for expert-parallel token dispatch, now the default for Qwen3 MoE and Mellum, removing the requirement that EP size equal TP size.

Adds `Trainer.loss_is_scaled_for_ga` attribute to declare whether `compute_loss` already scales loss for gradient accumulation.

Supports per-layer cache configuration so `DynamicCache` and `StaticCache` initialize each layer from its own config (sliding window, attention chunk size, conv states, and attention head counts) for heterogeneous models.

Adapts `Trainer` to work with expert parallelism.

Supports registering new processing backends.

NEW SUPPORTED PLATFORM EmbeddingGemma2 multimodal embedding model

NEW SUPPORTED PLATFORM XPU support for continuous batching

└──▷ BREAKING ON UPGRADE

! All MoE models whose routers compute logits now return them when `output_router_logits=True` via `MoeModelOutputWithPast` and the MoE causal LM output classes; code relying on the previous outputs or their absence must be updated.

! `Owlv2ForObjectDetection.embed_image_query` now selects the query box with the highest objectness score instead of the OWL-ViT heuristic, so image-guided query embeddings and detections may differ from earlier releases.

- ! The `'paged|'` prefix for SDPA and flash attention implementations is deprecated; use `sdpa` or `flash_attention_2` for continuous batching instead of `paged|sdpa` or `paged|flash_attention_2`.
- ! The regular `flash_attention.py` and `sdpa_attention.py` functions now support continuous batching directly, and `'paged|...'` implementations for these are redirected to them; eager still requires the `'paged|eager'` prefix.
- ! In continuous batching, the cache update for the index-based and block-table paths is fused into a single call, changing the cache update function's behavior and affecting any custom code that calls the separate update paths.

WAS THIS USEFUL? 👍 👎

Anyscale

Feedback ⓘ Version 0.26.109 release notes NOTES ↗ RANK

Anyscale provides a managed Ray platform for distributed AI development, training, batch processing, and model serving.

> Anyscale's 0.26.109 CLI/SDK release leads with scheduler config rollback, CA-bundle and proxy-aware networking, and an expanded job-logs flag, alongside a cluster of breaking changes around cloud registration, CLI flag parsing, and TLS trust — plus a docs update adding GPU pipelining for PyTorch batch inference.

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

01 Scheduler config rollback command NEW

90

Adds `anyscale scheduler config rollback --version <version>` CLI command and anyscale.scheduler.rollback_config(version=<version>) SDK method to roll back a scheduler config to a named prior version, which is applied as a new active version.

Roll back a scheduler config to a known-good prior version without being prompted for confirmation.

```
$ anyscale scheduler config rollback --version 4 --yes
```

— Exact CLI and SDK signatures plus runnable command Version 0.26.109 release notes

02 Kiro install target for anyscale skills NEW

90

Adds Kiro as an install target for `anyscale skills install` via `--platform kiro`; installs skills to `~/.kiro/skills` and hooks to `~/.kiro/hooks/anyscale-skills.json`; set `KIRO_HOME` to relocate both.

— Exact flag, install paths, and override variable named Version 0.26.109 release notes

03 Custom CA bundle trust for CLI and SDK BREAKING

85

Adds `ANYSCALE_SSL_CA_CERT` environment variable — set it to a CA bundle file path so the CLI and SDK trust that bundle for Anyscale API calls, enabling use behind TLS-inspecting proxies. Additionally, if `REQUESTS_CA_BUNDLE`, `CURL_CA_BUNDLE`, or `SSL_CERT_FILE` points to a file, the CLI and SDK now use that bundle instead of the default public CA bundle, which may affect environments relying on the old default.

Route Anyscale CLI/SDK traffic through a corporate TLS-inspecting proxy by pointing to your internal CA bundle.

```
$ export ANYSCALE_SSL_CA_CERT=/etc/ssl/certs/corporate-ca-
bundle.crt
export HTTPS_PROXY=http://proxy.corp.example.com:8080
anyscale job submit config.yaml
```

— Exact env vars and a runnable example given Version 0.26.109 release notes

04 **Fetch all cluster logs for a job** NEW 85

Adds `--all` flag to `anyscale logs job` to fetch every log file of the cluster that ran the job (worker logs, cluster startup output, etc.), equivalent to the glob `'*'`; without `--all`, only the driver log of the last attempt is returned.

Collect all cluster logs from a job run — including worker and startup logs — for post-mortem analysis.

```
$ anyscale logs job --id <job-id> --all
```

— Flag, default behavior, and example command all given Version 0.26.109 release notes

05 **Cloud registration and verification improvements** BREAKING 85

On AWS, `anyscale cloud verify`, `anyscale cloud register`, `anyscale cloud resource create`, and `anyscale cloud update --resources-file` now warn when the cross-account IAM role lacks the `ec2:DescribeCapacityReservations` action. `anyscale cloud register` prints the cloud ID and cloud resource ID before the Helm install command when registering a Kubernetes or KubeRay cloud, and includes the cloud ID in the success message for VM clouds. It also now requires object storage (`s3://<bucket>`, `gs://<bucket>`) for clouds on the Kubernetes compute stack and fails before registering anything when the bucket name is missing.

— Names every affected command and the new storage requirement

Version 0.26.109 release notes

06 **Proxy environment variable support for API and log calls** 80

BREAKING

The CLI and SDK now apply proxy environment variables (`no_proxy` / `NO_PROXY`, `HTTPS_PROXY`, proxy URL credentials, scheme-less proxy values) when calling the Anyscale API and downloading log files via `anyscale logs` and `anyscale job logs --follow`. On Python 3.9 and 3.10, log file downloads via `anyscale logs`

commands fail when the proxy URL starts with `https://`; use an `http://` proxy URL, add the storage bucket host to `no_proxy`, or upgrade to Python 3.11 or later.

— Names env vars and a Python-version regression with workarounds

Version 0.26.109 release notes

07 **`--working-dir` rejected with config file entrypoint** BREAKING

75

`anyscale job submit` now rejects `--working-dir` when a config file is passed as the positional entrypoint (e.g., `anyscale job submit config.yaml --working-dir ./src`); use `-f` or `--config-file` instead.

— Exact before/after command pattern and remedy given Version 0.26.109 release notes

08 **Ray Serve proxy placement via proxy_location** NEW 70

Adds `proxy_location` field to the service config to control where Ray Serve runs its proxies (e.g., `HeadOnly` to run a single proxy on the head node); `Disabled` is rejected.

— Config key and behavior named, no full example Version 0.26.109 release notes

09 **Empty-string support for --env across commands** IMPROVED

70

`--env` on `anyscale job submit`, `anyscale service deploy`, `anyscale workspace_v2 create`, and `anyscale workspace_v2 update` now accepts an empty value such as `--env FOO=`, setting the variable to an empty string.

— Flag and syntax given across four named commands Version 0.26.109 release notes

10 **Scheduler CLI visibility and output changes** BREAKING

60

The `anyscale scheduler` command group now appears in `anyscale --help` with commands such as `anyscale scheduler config apply` (in beta). The `-o` value for `anyscale scheduler config get` and `anyscale scheduler config list` is now case-sensitive (use `-o json`; `-o JSON` no longer works), and the list table adds a `CREATOR_ID` column.

— Names flags and columns but little mechanism beyond that Version 0.26.109 release notes

THINNER COVERAGE BELOW

11 **KubeRay connector status in cloud status output** NEW

55

Adds `connector_status` and `connector_status_details` fields to `anyscale cloud status` output for reporting Anyscale Connector status of a KubeRay cloud resource.

— Fields named but no example output shown Version 0.26.109 release notes

12 **GPU pipelining for PyTorch batch inference** NEW 45

Adds GPU pipelining support to overlap host-to-device copies with GPU compute for PyTorch batch inference via `TorchInference`.

— Names the API but gives no example or config detail product docs

└──▷ **BREAKING ON UPGRADE**

! The `-o` flag for `anyscale scheduler config get` and `anyscale scheduler config list` is now case-sensitive — `-o JSON` no longer works; use `-o json`.

! `anyscale job submit` now rejects `--working-dir` when a config file is passed as the positional entrypoint (e.g., `anyscale job submit config.yaml --working-dir ./src`); use `-f` or `--config-file` instead.

! If `REQUESTS_CA_BUNDLE`, `CURL_CA_BUNDLE`, or `SSL_CERT_FILE` points to a file, the CLI and SDK now use that bundle for Anyscale API calls instead of the default public CA bundle, which may affect environments that set these variables for other purposes.

! `anyscale cloud register` now requires object storage (`s3://<bucket>`, `gs://<bucket>`) for clouds on the Kubernetes compute stack and fails before registering anything when the bucket name is missing.

! On Python 3.9 and 3.10, log file downloads via `anyscale logs` commands fail when the proxy URL starts with `https://`; use an `http://` proxy URL, add the storage bucket host to `no_proxy`, or upgrade to Python 3.11 or later.

WAS THIS USEFUL? 👍 👎

Ollama

Feedback ⓘ v0.40.0 NOTES ↗ CODE ↗ RANK

Local large language model runner that downloads, serves, and chats with models through a CLI and REST API.

> Ollama v0.40.0 expands automatic MLX runtime execution on Apple Silicon to cover more model architectures, including decision models, with no extra configuration required.

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

01 **MLX runtime support for additional model architectures** NEW

55

Ollama now runs MLX-supported model architectures automatically on Apple Silicon without extra configuration, covering `gemma4`, `qwen3.5`, `qwen3.6`, and `qwen3.8`. MLX support was also added for decision models, including `nimble`, `tev1`, `clef`, and `clef-flash`.

— Names all model architectures but no config/flag or mechanism detail. v0.40.0

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

vMLX

Feedback ⓘ v1.6.75 NOTES ↗ RANK

The vMLX server runs compressed MLX models on Apple Silicon with disk caching, paged memory, continuous batching, and hybrid SSM scheduling.

> vMLX 1.6.75 focuses on DFlash2 drafter maturity — bundling it for Qwen3.8-27B, preserving its context across SSD snapshots, and surfacing its cache and usage metrics — alongside new model/kernel support and a stricter Responses API text format.

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

01 **text.format envelope in Responses API** NEW 65

Supports the `text.format` envelope in the Responses API and returns a strict failure when JSON output is invalid, rather than falsely marking the response as successful.

— Names exact API field and the before/after failure behavior. v1.6.75

02 **Bundled DFlash2 drafter for Qwen3.8-27B** NEW 60

Adds a bundled DFlash2 drafter for compatible Qwen3.8-27B bundles, loaded automatically but with explicit opt-out and manual drafter selection still available.

— Describes mechanism and opt-out but no exact flag named. v1.6.75

THINNER COVERAGE BELOW

03 **DFlash2 SSD snapshot context preservation** IMPROVED 50

DFlash2 SSD snapshots now preserve drafter context across prefix snapshots, enabling efficient warm decoding after restart.

— Explains benefit and mechanism but no config surface given. v1.6.75

04 **DFlash2 telemetry and usage reporting** NEW 50

Exposes DFlash2 cache and shared SSD pool telemetry in the app, reports prompt-cache reuse on SimpleEngine MLLM streams (DFlash2 session store) in usage output, adds copy-draft token counts as a separate field, and reports full prompt tokens in usage.

— Names several metrics surfaces but not exact API fields. v1.6.75

05 **Dense JANGH bundle support (Qwen3.8-27B JANGH2)** NEW

40

Loads dense JANGH bundles such as Qwen3.8-27B JANGH2, with or without DFlash2.

— Names bundle but gives no further operational detail. v1.6.75

06 **Qwen4 sparse AR decode and QSA kernels** NEW

40

Runs Qwen4 sparse AR decode and QSA score reduction kernels on MLX 0.32.3.

— Names kernels and MLX version but no usage detail. v1.6.75

07 **Copy drafts in native MTP verify cycle** IMPROVED

35

Adds copy drafts to the native MTP verify cycle while keeping wide verification exact.

— Thin mechanism description with no usage path. v1.6.75

08 **Mixed gate/up expert layouts in MoE path** NEW

30

Supports mixed gate/up expert layouts in the multi-row exact MoE path.

— Bare capability statement without scope or benefit. v1.6.75

WAS THIS USEFUL?  

DATA

◆ Vector DBs & RAG

Weaviate

Feedback ⓘ v1.40.0 NOTES ↗ RANK

Weaviate is an open-source vector database that stores both objects and vectors, allowing for the combination of vector search with structured filtering with the fault tolerance and scalability of a cloud-native database.

> Weaviate v1.40.0 ships GA Namespaces for multi-tenant isolation, a REST Search API with hybrid, near-object, near-vector and aggregate-counts endpoints enabled by default, deduplicated and RBAC-aware backups, a GA drop-vector-index endpoint, and 4-bit rotational quantization.

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

01 **Deduplicated backups with dedupeReplicas** NEW 90

Opt-in `dedupeReplicas` deduplicated backups archive converged shards once and fan the restore out to every replica, gated by a Weaviate license key, reducing storage use on multi-replica clusters.

Trigger a deduplicated backup that archives each converged shard once and fans the restore out to all replicas, saving storage on multi-replica clusters.

```
$ curl -X POST 'http://localhost:8080/v1/backups/s3' \
  -H 'Content-Type: application/json' \
  -d '{"id": "deduped-backup-2025", "dedupeReplicas": true}'
```

— Named flag, mechanism, and runnable curl example v1.40.0

02 **RBAC role data in backups via includeRoles** NEW 70

The backup/restore process adds an `includeRoles` option to include RBAC role data in backups so roles are automatically restored alongside data.

Back up a cluster including RBAC roles so they are automatically restored alongside data.

```
$ curl -X POST 'http://localhost:8080/v1/backups/s3' \
  -H 'Content-Type: application/json' \
  -d '{"id": "my-backup-2025", "includeRoles": true}'
```

— Named flag with a runnable curl example v1.40.0

03 **REST Search API hybrid, near-object, near-vector and aggregate endpoints**

NEW

68

The REST Search API adds `hybrid search`, `near-object`, `near-vector`, and `aggregate-counts` endpoints, plus `returnReferences` for structured reference selection, and is now default-on with no experimental flag required.

— Names every endpoint but gives no runnable example v1.40.0

04 **Drop-vector-index endpoint reaches GA** IMPROVED

60

The drop-vector-index endpoint is now enabled by default (GA), supporting RBAC and multi-tenancy, allowing reclamation of disk space by removing unused indices.

— Names endpoint and benefit but no usage example v1.40.0

THINNER COVERAGE BELOW

05 **Namespaces GA for multi-tenant cluster isolation** NEW

45

Namespaces reach GA, providing control-plane and data isolation between users on a shared cluster, gated by a Weaviate license key.

— Explains purpose but no config or command shown v1.40.0

06 **Alter Schema reindex property (Preview)** NEW

40

Introduces Alter Schema — Reindex Property (Preview) for changing a property's index types.

— Named but no mechanism or usage shown v1.40.0

07 **RQ4 4-bit rotational quantization** NEW

40

Adds RQ4 (4-bit Rotational Quantization) with centered quantization support.

— Named technique but no behaviour or usage detail v1.40.0

08 **HFresh routing, rerank budgets and usage reporting** IMPROVED

40

HFresh gains MUVERA decoupled routing and rerank budgets, plus HFresh postings and shared bucket counts added to the usage report.

— Names surfaces but no mechanism or example v1.40.0

09 **Replica movement management improvements** IMPROVED

40

Adds automated cleanup of stale replica movement operations and enables replica movement via runtime config.

— Names runtime config toggle but no exact key given v1.40.0

10 **Backup transport and coverage additions** IMPROVED

35

GCS backups now default to gRPC transport for `Backup-GCS`, and backups now also include GEO HNSW index data.

— Thin one-line additions with no mechanism detail `v1.40.0`

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

Feedback ⓘ snapshot-20261007 NOTES ↗

RANK

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

> LangSmith's latest snapshot tightens Model Access policy enforcement on the LLM Gateway, adds SSO deep-linking, IAM-based ECR access, and sandbox networking/resource controls, while capping extended trace retention at 180 days and renaming legacy role permissions.

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

01 **Redirect parameter for SSO login** NEW 85

Adds `redirect_to` query parameter to the organization SSO login URL (`/sso/login/<slug>?redirect_to=<path>`) so SAML sign-in lands on a specific page instead of the workspace root.

Deep-link a SAML user directly to a specific run after SSO sign-in, avoiding a second manual navigation step.

```
$ /sso/login/<slug>?
redirect_to=/o/<org>/projects/p/<project>/r/<run>
```

— Exact endpoint and runnable example provided snapshot-20261007

02 **Model Access policy controls for LLM Gateway** IMPROVED 83

`GET /v1/models` on the LLM Gateway now returns only the models permitted by the calling credential's model access policies. A new Model configurations provider option lets a workspace's saved model configurations (all or selected) be permitted, so requests resolving to a configuration are allowed rather than blocked. New Model Access policies also get a generated policy name that can be kept or edited before creation.

Retrieve only the models your credential is allowed to call, so downstream tooling never tries a model that will be rejected.

```
$ curl -H 'Authorization: Bearer <api_key>' https://<langsmith-
host>/v1/models
```

— Named endpoint and config options, setup flow not detailed snapshot-20261007

03 **180-day cap on extended trace retention** BREAKING 70

Extended retention for traces ingested into LangSmith SaaS is now capped at 180 days; new settings above 180 days are rejected and existing settings above 180 days are applied as 180 days for new traces. Self-hosted and BYOC deployments are unchanged.

— Clear before/after and scope, no migration command snapshot-20261007

04 **Renamed role permissions prompts and rules** BREAKING 65

Built-in workspace roles no longer return the retired `repos:*` and `run-rules:*` permissions, which have been renamed to `prompts:*` and `rules:*`; the old names are rejected when creating a custom role.

— Exact permission names given, no migration script snapshot-20261007

05 **Agent Auth connection visibility for admins** NEW 60

Workspace administrators can view workspace-level Agent Auth connections from the Integrations settings page, including authentication type and update details, and can inspect a connection's associated credential metadata without exposing secret or token material.

— UI path named, no API surface given snapshot-20261007

THINNER COVERAGE BELOW

06 **Model pricing tool in LangSmith MCP** NEW 50

Adds `list_model_price_maps` tool to LangSmith MCP for searching and paginating model pricing rules in a workspace.

— Named tool, no usage example given snapshot-20261007

07 **IAM role assumption for ECR access** NEW 50

LangSmith Cloud can now assume tagged customer IAM roles for ECR repository discovery and snapshot image pulls.

— Mechanism named, no setup steps given snapshot-20261007

08 **Raw TCP passthrough to allow-listed IPs** NEW 50

Sandboxes now pass raw TCP connections to literal public IP addresses when the destination IP and port are explicitly allow-listed.

— Mechanism named, no config example shown snapshot-20261007

09 **Context-aware LangSmith Chat in tracing projects** IMPROVED 50

LangSmith Chat in a tracing project now reads active filters, view, time window, and selection, and names runs and threads by what was asked rather than by the serving app.

— Explains behavior change, no config surface named `snapshot-20261007`

10 **Vertex AI models without stored service account key** NEW

50

Organizations with the LLM auth proxy enabled can select Google Vertex AI models in the playground and Agent Builder without storing a Vertex service account key in the workspace.

— Names feature and UI areas, no setup steps `snapshot-20261007`

11 **Larger Context Hub repository capacity** IMPROVED

40

Context Hub repositories now support up to 1,500 files, allowing agents with larger skill collections to sync successfully.

— Concrete limit, no usage detail `snapshot-20261007`

12 **Higher open-file limit in sandboxes** IMPROVED

40

Sandboxes now use the guest kernel's maximum open-file limit, allowing highly concurrent agents and subprocesses to run without exhausting the default descriptor ceiling.

— Mechanism named, no numeric limit given `snapshot-20261007`

13 **New revision on deployment command updates** IMPROVED

40

Updating install or build commands through the deployment API now creates a new revision.

— Names API behavior, no endpoint path given `snapshot-20261007`

14 **Contract expiration warnings for admins** NEW

25

LangSmith Cloud warns organization admins before an enterprise contract or cancelled plan expires.

— Bare description, no timing or UI detail `snapshot-20261007`

└──▷ **BREAKING ON UPGRADE**

! Extended retention settings above 180 days on LangSmith SaaS are applied as 180 days for new traces; new settings above 180 days are rejected outright. Self-hosted and BYOC deployments are unchanged.

! Built-in workspace roles no longer return the retired `repos:*` and `run-rules:*` permissions (renamed to `prompts:*` and `rules:*`); these are rejected when creating a custom role.

WAS THIS USEFUL? 👍 👎

MLflow

Feedback ⓘ v3.17.0 NOTES ↗ RANK

MLflow tracks, evaluates, and monitors machine learning and LLM applications, covering experiment tracking, tracing, prompt management, and a model registry.

> MLflow shipped a one-line setup wizard for bootstrapping tracing, programmatic experiment/run restoration, opt-in SQL trace analytics rollups that require a mandatory database schema upgrade, fine-grained sub-resource RBAC, and expanded AI Gateway and TypeSafe judge support in v3.17.0.

└──▷ **WHAT SHIPPED · 18 FEATURES** most completely described first ? what's the number?

01 **SQL trace analytics daily rollups** BREAKING 95

Introduces opt-in SQL trace analytics daily rollups to reduce query overhead for large trace datasets, enabled via `mlflow db upgrade <database-url>`. SQL-backed servers require this schema upgrade before restarting on v3.17.0: stop all writers, take a restorable backup, run `mlflow db upgrade <database-url>`, then restart all replicas — mixed-version rolling upgrades are not supported.

Upgrade the database schema (required before restarting replicas on v3.17.0) to enable trace analytics daily rollups.

```
$ mlflow db upgrade postgresql://user:pass@db-host:5432/mlflow
```

— Exact migration command with full upgrade runbook. v3.17.0

02 **One-line tracing setup wizard** NEW 80

A new setup wizard, run via `curl -LsSf https://mlflow.org/wizard/setup.sh | sh`, connects a project to an MLflow server or Databricks workspace and launches a coding agent (Claude Code, Codex, or OpenCode) to add tracing automatically.

Bootstrap MLflow tracing in a new project without manual configuration — the wizard connects to your server and lets a coding agent instrument your code.

```
$ curl -LsSf https://mlflow.org/wizard/setup.sh | sh
```

— Runnable one-liner, though agent behaviour is opaque. product docs

03 **Experiment and run restoration API** NEW 80

Adds mlflow.restore_experiment() and mlflow.restore_run() to the fluent API for programmatically restoring a soft-deleted experiment or run back to active state.

Restore a soft-deleted experiment and a specific run back into active state without leaving Python.

```python
import mlflow

mlflow.set_tracking_uri('http://localhost:5000')

# Restore a deleted experiment by ID
mlflow.restore_experiment('3')

# Restore a deleted run by run ID
mlflow.restore_run('a1b2c3d4e5f6a1b2c3d4e5f6a1b2c3d4')
```

— Exact API calls with a runnable example. v3.17.0

04 **TypeSafe and OpenRouter judge support in Gateway** NEW

75

Gateway System One now supports TypeSafe and OpenRouter judges; direct TypeSafe judges can use a custom `base_url`, and TypeSafe Gateway endpoints configure via `typesafe_api_base`. TypeSafe model support was also added to built-in scorers and custom judges, returning typed Boolean or categorical feedback.

— Names the two config keys but gives no example. v3.17.0

05 **AI Gateway disable switch and model discovery** NEW

60

Adds `MLFLOW_ENABLE_AI_GATEWAY` environment variable allowing administrators to disable the AI Gateway, OpenAI-compatible model discovery to the Gateway, and Anthropic Messages passthrough to `AmazonBedrockProvider`.

— Names env var and provider class, no example. v3.17.0

THINNER COVERAGE BELOW

06 **Logged model artifacts in presigned upload endpoint** IMPROVED

55

Supports logged model artifacts in the `createPresignedUploadUrl` endpoint.

— Named endpoint but no request example. v3.17.0

07 **Span description support** IMPROVED

55

Adds first-class span description support to `@mlflow.trace` and `mlflow.start_span`.

— Names both decorators, no example call. v3.17.0

08 **Fine-grained sub-resource RBAC** NEW 50

Adds fine-grained sub-resource permissions for runs, traces, and versions with wildcard grants and explicit DENY.

— Describes scope but no config syntax shown. v3.17.0

09 **order_by in search_prompts()** IMPROVED 50

Supports `order_by` in search_prompts() across the client, fluent, and GenAI APIs.

— Named parameter, no call example. v3.17.0

10 **UC model service dependency support** NEW 50

Adds `uc_model_service` UC model-version dependency support and a `DatabricksUCModelService` resource type.

— Named resource types, no configuration example. v3.17.0

11 **Evaluation-dataset name IN filter** IMPROVED 50

Supports name IN (...) filters for evaluation-dataset search.

— Named filter syntax, no sample query. v3.17.0

12 **TypeSafe AI tracing and autologging** NEW 45

Adds `@mlflow/typesafe` tracing integration and autologging support for TypeSafe AI.

— Named integration, no usage example given. v3.17.0

13 **Helm chart operational improvements** IMPROVED 45

Exposes Helm liveness/readiness/startup probes as configuration variables, publishes Helm charts after stable release images, and configures topology spread constraints in the Helm chart.

— Names three Helm changes without exact keys. v3.17.0

14 **Prometheus OTLP ingestion metrics** NEW 45

Exposes OTLP ingestion metrics for Prometheus via the tracing/tracking layer.

— Names the metrics surface, no endpoint or example. v3.17.0

15 **TraceImage component for custom trace views** NEW 35

Adds a `TraceImage` component to custom trace views.

— Bare component name, no usage shown. v3.17.0

16 **Client-side LLM trace cost computation for SageMaker** NEW 35

Computes LLM trace costs client-side for SageMaker tracking.

— Bare description, no mechanism given. v3.17.0

---

17 **UI additions for model version and dataset ID** IMPROVED

35

Adds a source logged model link to the model version page in the UI, and a dataset ID column to the UI.

— Two thin UI additions, no screenshots provided. v3.17.0

---

18 **Source metadata in TypeScript agent traces** IMPROVED

25

Adds source metadata to TypeScript agent traces.

— Minimal description, no field names listed. v3.17.0

└──▷ **BREAKING ON UPGRADE**

! SQL-backed servers require a schema upgrade before restarting on v3.17.0: stop all writers, take a restorable backup, run `mlflow db upgrade <database-url>`, then restart all replicas. Mixed-version rolling upgrades are not supported.

WAS THIS USEFUL? 👍 👎

UK AI Security Institute Inspect AI

Feedback ⓘ

0.3.277 CODE ↗ RANK

Inspect is the UK AI Security Institute's framework for large language model evaluations, covering datasets, solvers, scorers, and agentic eval tasks.

> Inspect AI shipped accurate cost tracking that prices requests at the rates of the model that actually served them, a new `epoch=N` parameter for repeating samples via enqueue_sample(), and an `inspect sandbox cleanup docker` subcommand, alongside tighter Agent Bridge configuration enforcement and several breaking changes including a new `openai >= 3.4.0` requirement.

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

01 **Repeated sample epochs via enqueue_sample** NEW 80

Adds an `epoch=N` parameter to enqueue_sample() so a custom sample source can run the same sample repeatedly under its own id, labelled as a specific epoch, without duplicating the sample definition.

Run the same adversarial sample 10 times in a row under its own ID, labelled as epoch 3, without duplicating sample definitions.

```python
enqueue_sample(samples, epoch=3)
```

Run a single sample repeatedly as distinct epochs from a custom source — useful for measuring variance across repeated model calls on the same input.

```python
enqueue_sample(samples, epoch=3)
```

— Runnable example and exact parameter name given 0.3.277

02 **Docker sandbox cleanup subcommand** NEW 80

Adds an `inspect sandbox cleanup docker <project>` subcommand that removes custom networks declared in a compose file and cleans up older interrupted logs and sample buffers.

Clean up a Docker sandbox project including custom networks and leftover interrupted logs after a cancelled evaluation run.

```
$ inspect sandbox cleanup docker <project>
```

— Exact runnable subcommand with scope of cleanup given `product docs`

03 **Cost tracking by actual serving model** NEW 75

Adds ModelAPI.served_model_usage() so cost tracking prices a request at the rates of the model that actually served it, covering refusal fallbacks, Azure deployments, and routers on OpenRouter, Bedrock, Fireworks and LiteLLM proxies.

— Names method and every covered provider/routing case `product docs`

04 **Agent Bridge follows eval configuration** IMPROVED 75

Agent Bridge requests are now governed by the eval's configuration: provider tool options such as web search domains, service tier, storage, and truncation all follow the eval config, and the sandbox agent bridge serves unknown model names with the eval's model; requests carrying `previous_response_id` are refused.

— Names every governed option and the refusal condition `0.3.277`

05 **Sandbox poll timeout recovery** NEW 65

Adds an opt-in `poll_timeout_recovery` parameter to sandbox tools so that an `exec_remote` poll that times out during a temporary sandbox stall keeps polling the running command instead of failing the call.

— Names the parameter and affected call but not failure modes handled `0.3.277`

06 **eval_config.token_limit_type preserved** IMPROVED 65

Preserves `eval_config.token_limit_type` from an exported or handwritten run config as the token-limit metering type, instead of rejecting it as an unknown generate option.

— Names the config key and prior failure it fixes `0.3.277`

07 **Images in tool results follow log_images setting** IMPROVED 60

Images in tool results recorded in the transcript now follow the `log_images` setting — stored as attachments when enabled and removed when disabled — matching how images in messages are already handled.

— Names the config setting and its two behaviours `0.3.277`

THINNER COVERAGE BELOW

08 **ModelEvent records requested model name** IMPROVED 55

Records the model name a bridged client requested on `ModelEvent` (field, hook, schema, and dataframe column).

— Names the field and surfaces it touches `0.3.277`

09 **Model API request and response id logging** IMPROVED

50

Model API events now log provider request ids (including for retried requests) and response ids, even when raw model API calls are not logged.

— Clear behaviour change but no config surface to act on `product docs`

10 **Model calls refused when limit already reached** IMPROVED

50

A model call is now refused before it is sent when a token or cost limit is already reached, including when usage exactly equals the limit.

— Describes exact trigger condition but no config name `product docs`

11 **Unknown Meta models default to frontier specs** IMPROVED

50

Unknown Meta models not yet in the model database now use the current frontier model's context window and `max` reasoning effort support.

— Names the fallback behaviour and reasoning-effort value `0.3.277`

12 **OpenAI dependency raised to 3.4.0** BREAKING

50

The OpenAI providers, message and output converters, and agent bridge now require `openai >= 3.4.0`.

— Exact version requirement given, migration path implied `product docs`

13 **Mistral image URLs recorded as text** BREAKING

50

Mistral: image URLs in model output are no longer downloaded from any host; they are instead recorded as text with the URL.

— Clear before/after behaviour described `product docs`

14 **Batch mode separates requests by HTTP headers** IMPROVED

45

In batch mode, requests with different HTTP headers are now sent in separate batches, so one request's headers no longer apply to others.

— Clear before/after but no named flag `product docs`

15 **cache_path rejects unsafe model names** BREAKING 45

cache_path(model) rejects unsafe model names with an error, and unsafe model names are not cached.

— Names the function and new error behaviour product docs

16 **Core data types moved to inspect_ai.core** IMPROVED 40

Moves core data types to the new `inspect_ai.core` package; existing import paths continue to work.

— States the move but no detail on what moved product docs

17 **shuffle accepts integer seed** IMPROVED 40

Dataset loaders' `shuffle` option now accepts an integer seed.

— Names the option but no usage example 0.3.277

18 **Wire-type modules load independently** IMPROVED 40

Wire-type modules now load without requiring the rest of `inspect_ai`.

— States the change with no mechanism detail 0.3.277

19 **Remote MCP servers refused under approval policy** BREAKING 40

Remote MCP servers are now refused while an approval policy is active.

— States the restriction with no further mechanism product docs

20 **ModelOutput.from_message optional parameter** IMPROVED 35

ModelOutput.from_message() gains support for an additional optional parameter.

— Named method but parameter itself unspecified product docs

└──▷ **BREAKING ON UPGRADE**

! The OpenAI providers, message and output converters, and agent bridge now require `openai >= 3.4.0`.

! Mistral: Image URLs in model output are no longer downloaded from any host; they are recorded as text with the URL.

! Remote MCP servers are now refused while an approval policy is active.

! Agent Bridge: requests with `previous_response_id` are refused.

! cache_path(model) rejects unsafe model names with an error; unsafe model names are not cached.

WAS THIS USEFUL?  

Braintrust

Feedback ⓘ snapshot-20261007 NOTES ↗ RANK

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

> Braintrust's biggest changes this window were versioned saves for custom views, expanded audit logging for project groups and scores, new sub-agent cost/token metadata surfaced in a Coding Agent Insights dashboard, and new integrations (Roast AI, pi) alongside a Ruby SDK release with trace redaction and a project-level AI provider credentials API.

└──▷ **WHAT SHIPPED · 13 FEATURES** most completely described first ? what's the number?

01 Versioned saves for custom views NEW

88

Adds versioned save options via the Save changes action in the Edit view editor: Save as new version (adds a new version of the view), Save as new view (creates a separate view leaving the original unchanged), and Revert changes (discards unsaved edits). Also adds a Save to project confirmation prompt when closing the editor with unsaved changes, and the ability to share a new custom view with the team via the Save view action.

Save an updated custom view as a new version so teammates see your changes while the previous version remains in history.

> 📍 In the Braintrust UI, open the custom view and click Edit view. Make your changes, then select Save changes. In the Update saved view dialog, choose Save as new version.

Experiment with a custom view layout without overwriting the shared original — save your edits as a separate view.

> 📍 In the Braintrust UI, open the custom view and click Edit view. Make your changes, then select Save changes. In the Update saved view dialog, choose Save as new view.

— Names every save/share action and dialog exactly product docs

02 Audit logs for project groups and scores NEW

75

Adds `project_group`, `project_group.created`, `project_group.updated`, and `project_group.deleted` audit-log event types for project group lifecycle changes, and `project_group_project`, `project_group_project.created`, `project_group_project.deleted` for project group membership changes. Adds `project_score`, `project_score.created`, `project_score.updated`, `project_score.deleted` event types for project score configuration changes (excluding scoring runs and human review score reordering); these were later extended

to also cover online scoring rules and human review score configurations, including pause and resume status changes.

— Lists every audit event type across both releases `product docs`

---

03 **Sub-agent token usage and cost tracking in spans**

`UPDATED INTEGRATION` `NEW` 75

Agent run spans now include sub-agent token usage across all models and `total_cost_usd` metadata, viewable in the built-in Coding Agent Insights dashboard broken down by user, API key, model, and branch, including expandable per-day Loop automation findings that flag recurring inefficiencies.

Inspect coding-agent cost and usage broken down by user, model, and branch in the built-in Braintrust dashboard.

📍 1. In the Braintrust console, open the Coding Agent Insights dashboard. 2. Review the `Usage by user`, `Usage by API key`, `Usage by model`, and `Usage by branch` panels. 3. Expand a day's Loop automation findings to see recurring inefficiencies flagged across team sessions.

— Names metadata field and exact dashboard panels `product docs`

---

04 **Ruby SDK span customizers for trace redaction** `NEW`

75

Adds Ruby SDK span customizers for redacting sensitive data from traces before they reach Braintrust, documented at `/docs/sdks/ruby/redact-sensitive-data`.

Redact a sensitive header and attach an environment tag to every span before export, using the new `span_customizers:` hook in `Braintrust.init`.

text

```
Braintrust.init(
  project: 'my-agent',
  span_customizers: [
    lambda { |span|
      span.metadata['env'] = 'production'
      span.input&.delete('authorization')
    }
  ]
)
```

— Includes doc path and runnable redaction example `product docs`

---

05 **Project-level AI provider credentials API** `NEW` 70

Adds `GET` / `POST` project-level AI provider credentials management via API, enabling programmatic creation and update of per-project provider configuration.

— Names HTTP methods but omits exact endpoint path product docs

06 **Typed span context manager and span processor params** IMPROVED

60

Adds `with start_span(...) as span` type-checking as `Span` under pyright and mypy, and fully typed add_braintrust_span_processor() parameters.

— Names exact function and type-checking tools product docs

THINNER COVERAGE BELOW

07 **Chart data export from dashboards** NEW

55

Dashboard chart export now covers SQL charts and table charts, and chart data-point click menus include an `Export data` option that downloads JSON or CSV.

— Names chart types and export formats product docs

08 **Custom preprocessors fail runs on error** BREAKING

55

Custom preprocessors that throw an error on any span now fail the entire run instead of silently skipping that span; existing preprocessors with error-prone logic will begin failing runs on upgrade.

— States exact before/after behavior change product docs

09 **pi coding session tracing and MCP server access** NEW

50

Adds integration with pi coding sessions, enabling traces to Braintrust and use of the Braintrust MCP server to query projects, experiments, logs, and datasets from pi.

— Names MCP server capability, no command given product docs

10 **Roast AI workflow tracing integration** NEW INTEGRATION NEW

45

Adds integration with Roast AI to trace workflows and individual cogs in Braintrust for step-by-step execution debugging, LLM call inspection, and workflow failure monitoring.

— Describes integration scope, no setup steps given product docs

11 **Thread panel beside trace view** IMPROVED

40

On desktop, the thread now opens in a panel beside the trace so conversation and spans are visible side by side.

— Describes UI layout change, lacks further detail product docs

12 **Immediate timeout for zero-value signal timeout** BREAKING

35

A `signal` `timeout` of `0` now times out immediately.

— Single-line behavior change, no further context product docs

13 **SDK releases: TypeScript, Python, Ruby** NEW

35

Adds Braintrust TypeScript SDK `v3.37.0` API reference, Braintrust Python SDK `v0.44.1` API reference, and a Ruby SDK `v0.6.0` release.

— Version numbers only, no changelog detail given product docs

NEW INTEGRATION Eve 0.62+ tracing (TypeScript SDK)

UPDATED INTEGRATION AgentScope TeamPipeline.reply_stream() tracing (Python SDK)

UPDATED INTEGRATION Anthropic MCP connector tool call spans (Python SDK)

└──▷ **BREAKING ON UPGRADE**

! Custom preprocessors that throw an error on any span now fail the entire run instead of silently skipping that span — existing preprocessors with error-prone logic will begin failing runs on upgrade.

! A `signal` `timeout` of `0` now times out immediately.

snapshot-20261007 NOTES ↗ LATEST STABLE

Adds Coding Agent Insights dashboard, SQL charts, Loop on dashboards, DPAPI credential storage on Windows, new thread view, and expanded span customizer coverage across Python, TypeScript, Go, and Ruby SDKs.

Coding Agent Insights dashboard (public preview, requires data plane v2.13.0+) shows usage by user, API key, model, tool, and branch with cost breakdowns, plus a daily Loop automation for surfacing inefficiencies.

New thread view (public preview) groups trace turns into request, agent work, and response, with tool calls and sub-agent work nested and Debugger findings highlighted inline.

SQL chart builder (public preview, Pro/Enterprise, requires data plane v2.1.1+) lets you write SQL or use a visual builder and plot results as time series, top lists, or tables across any dashboard.

Loop is now available from individual dashboards and the chart editor so you can build or refine charts in natural language with the dashboard as context.

`bt` CLI on Windows stores credentials in a DPAPI-encrypted file at `%USERPROFILE%\.braintrust\auth\credentials.dpapi`, shared between terminal and MSIX desktop app sessions.

+5 more

└─▷ **BREAKING ON UPGRADE**

! After downgrading from a `bt` version that uses `credentials.dpapi` to an earlier version on Windows, you must log in again — saved credentials are not back-migrated.

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

OpenAI

Feedback ⓘ changelog · Oct 7, 2026 CHANGELOG ↗ RANK

OpenAI provides AI models and APIs for building applications that generate and reason with text, images, and audio.

> OpenAI's latest changelog introduces the beta Decisions API, an Ultrafast response mode, the GPT-6.1 Sol model with multi-agent support, a browser tool for Agents, HIPAA compliance onboarding, and a simplified three-tier usage structure.

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

01 **Ultrafast service tier for GPT-6 Astra** NEW 80

Adds `service_tier: "ultrafast"` to the Responses API for the `gpt-6-astra` model to reduce inter-token latency; available globally with US data residency, but EU regional inference is not supported.

Enable Ultrafast mode on a GPT-6 Astra Responses API call to minimize time-between-tokens in latency-sensitive pipelines such as real-time voice or streaming UI.

```
$ curl https://api.openai.com/v1/responses \
  -H 'Authorization: Bearer $OPENAI_API_KEY' \
  -H 'Content-Type: application/json' \
  -d '{"model": "gpt-6-astra", "service_tier": "ultrafast",
"input": "Summarise the following log file in three bullet
points: ...'"}'
```

— Runnable curl example with exact flag and constraints changelog-20261007-ae983687

02 **Decisions API (beta)** NEW 70

Adds the `v1/decisions` endpoint, which turns text and images into typed answers and is claimed to be 10x faster than the Responses API.

— Endpoint named but no example call shown changelog-20261007-ae983687

03 **GPT-6.1 Sol model launch** NEW 70

Adds `gpt-6.1-sol`, a new model targeting complex coding and professional work at lower cost than GPT-6 Astra, priced at $2/1M input tokens, $0.10 cached input, $2.50 cache write, and $10 output, supporting up to 272K input tokens.

— Full pricing and context limit but no call example changelog-20261007-ae983687

04 **HIPAA compliance support flow** NEW 60

Adds an in-product flow under `Organization settings > General` letting eligible org admins review eligibility and accept a Business Associate Agreement (BAA) to enable HIPAA compliance for the organization.

Accept the BAA and enable HIPAA compliance support so your organization can process protected health information through the API.

📍 1. Sign in to `platform.openai.com` as an organization admin. 2. Go to Organization settings › General. 3. Locate the HIPAA compliance support section. 4. Review eligibility requirements, then click Accept BAA to enable HIPAA compliance support for your organization.

— Exact UI navigation path given, but no API detail changelog-20261007-ae983687

THINNER COVERAGE BELOW

05 **Browser tool for Agents API** NEW 45

Adds a browser tool to the Agents API enabling agents to complete tasks in an OpenAI-hosted browser, with website access approvals and sign-in handled by the calling application.

— Describes mechanism but no flag or endpoint named changelog-20261007-ae983687

06 **Multi-agent support in GPT-6.1 Sol** NEW 40

Adds multi-agent support (beta) to GPT-6.1 Sol via the Responses API, allowing the model to delegate work to subagents.

— Mechanism named but no usage example or config changelog-20261007-ae983687

07 **Simplified API usage tiers** IMPROVED 40

Simplifies API usage tiers from five to three — Build, Launch, and Grow — with automatic upgrades as total credit purchases reach tier minimums.

— Tier names given but no thresholds or benefits listed changelog-20261007-ae983687

WAS THIS USEFUL? 👍 👎

Mistral AI

Feedback ⓘ snapshot-20261007 NOTES ↗ RANK

Learn how to deploy and use Mistral AI's Large Language Models with our comprehensive documentation, guides, and tutorials.

> Mistral AI shipped its largest documented update window: Mistral Large 4 (1M context, MoE, multimodal) enters Public Preview alongside Mistral Medium 3.5 and Mistral Small 4, OCR 4.1 reaches General Availability with new block/word-level confidence scoring and page-range selection, and Agents/Conversations gain Custom Guardrails. These sit on top of dozens of prior additions spanning Voxtral audio models, Devstral/Magistral/Codestral/Ministral model families, new Agents/Batch/Moderation/Fine-tuning/Document Library APIs, Completion API parameters, and several deprecations.

└──▷ **WHAT SHIPPED · 13 FEATURES** most completely described first ? what's the number?

01 **OCR API parameters and model upgrades** IMPROVED 90

`mistral-ocr-4-1` (OCR 4.1) is now Generally Available and `mistral-ocr-latest` points to it. The OCR API gains `confidence_scores_granularity` (`'page'` for page-level scores, `'block'` for page- and block-level scores, `'word'` for page- and word-level scores), `include_blocks` (returns a `blocks` array per page with paragraph-level bounding boxes and structural labels in reading order), a `pages` parameter accepting comma-separated digits and ranges (e.g. `'0,1,2'`, `'0-5'`, `'0,2-4'`) in addition to a list of integers, `table_format` (choice of `markdown` or `html`), and `extract_footer` / `extract_header` parameters plus hyperlinks output. Earlier releases added `mistral-ocr-2512` (OCR 3) with the same table/footer/header/hyperlink options, `mistral-ocr-2505` (OCR 2) with annotations support, and `mistral-ocr-2503` (OCR) for document understanding.

— Enumerates every OCR parameter and value verbatim snapshot-20261007

02 **Deprecated models and stricter API enforcement** DEPRECATED 90

`mistral-ocr-4-0` is deprecated and retires on September 30, 2026 (migrate to `mistral-ocr-4-1` or `mistral-ocr-latest`); `labs-leanstral-1-5` (Leanstral 1.5) is deprecated and retires on September 30, 2026; `zai-glm-5-2` (Z.ai GLM 5.2) is deprecated and retires on October 31, 2026 (migrate to `zai-glm-5-3`); and Devstral 2.0 moves to paid API access. The API now strictly enforces validity of all parameters — requests with unsupported parameters that were previously silently ignored now return the error `'Extra inputs are not permitted'`. Documentation paths previously under `/le-chat/*` and `/mistral-vibe/*` now redirect to `/vibe/*`.

— Gives exact retirement dates and migration targets snapshot-20261007

03 **Voxtral audio models and Transcriptions API** NEW 80

Adds `voxtral-tts-2603` (Voxtral TTS): a text-to-speech model with zero-shot voice cloning, multilingual support, and real-time streaming. Adds `voxtral-mini-2602` (Voxtral Mini Transcribe 2) and `voxtral-mini-transcribe-realtime-2602` (Voxtral Mini Transcribe Realtime) with context biasing and `diarize` support in the Audio Transcriptions API. Earlier releases added `voxtral-small-2507` (Voxtral Small) for chat use cases, `voxtral-mini-2507` (Voxtral Mini), and `voxtral-mini-transcribe` (Voxtral Mini Transcribe) via the `audio/transcriptions` endpoint optimized for transcription, along with SDK support for Audio and Transcription.

— Names every audio model id and the transcriptions endpoint snapshot-20261007

04 **Mistral Medium 3.5 reasoning model** NEW 70

Adds `mistral-medium-3-5` in Public Preview: a frontier-class multimodal model optimized for agentic and coding use cases with adjustable reasoning via the `reasoning_effort` parameter, released as open weights under a Modified MIT license.

— Names tunable parameter and license but no usage walkthrough snapshot-20261007

05 **Custom Guardrails for Agents and Conversations** NEW 70

Guardrails can now be configured directly on an Agent via the `guardrails` parameter, or passed per-request on `POST /v1/conversations`.

— Names exact parameter and endpoint but not guardrail types snapshot-20261007

06 **Chat Completions API parameter additions** IMPROVED 70

Adds `presence_penalty` and `frequency_penalty` parameters to the chat completions API, an `n` parameter (number of completions per request, with input tokens billed only once), token usage information in streaming requests (available in the last chunk returned), and JSON mode for all models on La Plateforme.

— Names each parameter and its billing/streaming behavior snapshot-20261007

07 **Streaming security field and mandatory tool_call_id** BREAKING 70

Adds a `p` parameter to chunks streamed back by the Completion API as a security measure against token-length side-channel attacks; applications using the official SDK are unaffected, but users relying on `mistral-common` may need to update to 1.8.4 or higher. `tool_call_id` is now a mandatory field in chat messages with the `tool` role for function calling.

— Gives exact SDK version threshold and affected field snapshot-20261007

08 **New language, code, and reasoning models** NEW 65

Adds `labs-leanstral-2603` (Leanstral), an open-source code agent for Lean 4 formal proof engineering, and `mistral-moderation-2603` (Mistral Moderation 2603). Adds `magistral-medium-2509` / `magistral-small-2509` (Magistral 1.2), `codestral-2508` (Codestral 2508), `magistral-medium-2507` / `magistral-small-2507` (Magistral 1.1), `devstral-small-2507` / `devstral-medium-2507` (Devstral 1.1), `codestral-embed` (Codestral Embed), `devstral-small-2505` (Devstral Small), `mistral-small-2503` (Mistral Small 3.1), `mistral-saba-2502` (Mistral Saba), `mistral-small-2501` (Mistral Small 3), `codestral-2501` (Codestral 2501), `mistral-large-2411` / `pixtral-large-2411` (Mistral Large 2.1 / Pixtral Large), `ministral-3b-2410` / `ministral-8b-2410` (Ministral 3B/8B), `pixtral-12b-2409` / `mistral-small-2409` (Pixtral / Mistral Small v24.09), `mistral-large-2407` (Mistral Large 2), `mistral-small-2506` / `magistral-medium-2506` / `magistral-small-2506` (Mistral Small 3.2, Magistral Medium/Small), `labs-mistral-small-creative` (Mistral Small Creative, Labs), `devstral-2512` / `labs-devstral-small-2512` (Devstral 2 / Devstral Small 2), `mistral-large-2512` (Mistral Large 3) with `ministral-3b-2512` / `ministral-8b-2512` / `ministral-14b-2512` (Ministral 3), `mistral-medium-2505` (Mistral Medium 3), and `labs-leanstral-1-5` (Leanstral 1.5, an updated Lean 4 model with improved SFT mixture quality and extended long-context reasoning).

— Lists every model id but gives little behavioral detail per model `snapshot-20261007`

09 **Mistral Large 4 flagship model** NEW 60

Adds `mistral-large-4` (Mistral Large 4) in Public Preview: an open-weight, general-purpose multimodal model built on a Mixture-of-Experts architecture with a 1M token context window.

— Names architecture and context but no usage example `snapshot-20261007`

10 **Mistral Small 4 hybrid model** NEW 60

Adds `mistral-small-2603` (Mistral Small 4): a hybrid model unifying instruct, reasoning, and coding capabilities in a single multimodal model with a 256k token context window.

— States capability and context size, no further mechanism `snapshot-20261007`

THINNER COVERAGE BELOW

11 **New platform APIs: Agents, Batch, Moderation, Fine-tuning, Document Library, Classifier Factory**

NEW 55

Adds the Agents API (initial July 2024 release) and later a new, redesigned Agents API. Adds a moderation API and a batch API, plus inline batching that enables creating batch jobs without file uploading. Adds a fine-tuning API with support for `codestral-`

`latest`, Mistral Nemo, and Mistral Large. Adds a Document Library API to manage libraries. Adds a Classifier Factory. Adds custom structured outputs for all models.

— Names each API surface but not endpoints or schemas `snapshot-20261007`

---

12 **Vibe unified agent with Work, Code, and Chat modes** NEW

55

Adds Vibe, a unified agent at `chat.mistral.ai` with Work, Code, and Chat modes, including a Vibe CLI and a VS Code extension.

— Names the product surface and modes but no workflow detail `snapshot-20261007`

---

13 **Platform billing and enterprise administration** NEW

40

Adds a free API tier on La Plateforme, multiple currency support in the payment system including the option to pay in US dollars, and enterprise platform features including admin management.

— Thin prose descriptions of billing and admin features `snapshot-20261007`

└──▷ BREAKING ON UPGRADE

! A new parameter `p` is added to chunks streamed by the Completion API; applications relying on strict parsing of chunks may break. Applications using the official SDK are unaffected, but users relying on the `mistral-common` package may need to update to 1.8.4 or higher.

! The API now strictly enforces validity of all parameters; previously unsupported parameters were silently ignored — requests with unsupported parameters now return the error `Extra inputs are not permitted`.

! `tool_call_id` is now mandatory in chat messages with the `tool` role for function calling.

! `mistral-ocr-4-0` is deprecated and retires on September 30, 2026; migrate to `mistral-ocr-4-1` or `mistral-ocr-latest`.

! `labs-leanstral-1-5` (Leanstral 1.5) is deprecated and retires on September 30, 2026.

! `zai-glm-5-2` (Z.ai GLM 5.2) is deprecated and retires on October 31, 2026; migrate to `zai-glm-5-3`.

! Devstral 2.0 moves to paid API access.

! Previous `/le-chat/*` and `/mistral-vibe/*` documentation paths now redirect to `/vibe/*`.

WAS THIS USEFUL? 👍 👎

Anthropic

Feedback ⓘ snapshot-20261007 NOTES ↗ RANK

Anthropic provides Claude AI models and APIs for building applications that generate, analyze, and automate work.

> Anthropic promoted its Admin API organization-management methods out of beta across all seven SDKs and the `ant` CLI, added a Models API field to discover thinking-disable support, documented new cross-model thinking-block compatibility and refusal behavior, and deprecated Claude Sonnet 4.5 ahead of its November 30, 2026 retirement.

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

01 Admin API organization management across SDKs and CLI

IMPROVED 95

Admin API methods for organizations graduated out of beta and are now callable under `client.organization` (SDKs) and `ant organization` (CLI): client.organization.retrieve(); client.organization.users.list(), client.organization.users.update(), client.organization.users.remove(); client.organization.invites.create(), client.organization.invites.list(), client.organization.invites.delete(); client.organization.workspaces.members.add(), client.organization.workspaces.members.list(), client.organization.workspaces.members.update(), client.organization.workspaces.members.remove(); and client.organization.api_keys.list(), client.organization.api_keys.update(). The promotion also covers organization info, members, invites, workspaces, API keys, rate limits, service accounts, workload identity federation, customer-managed encryption keys, and compliance settings, and ships in Python SDK 1.10.0, TypeScript SDK 0.130.0, C# SDK 12.52.0, Go SDK 1.77.0, Java SDK 2.67.0, PHP SDK 0.53.0, Ruby SDK 1.75.0, and `ant` CLI 1.37.0; `client.beta.organization` and `ant beta:organization` remain available as aliases.

— Names every method, endpoint area, and exact SDK version snapshot-20261007

02 Thinking support discovery in Models API NEW

88

The Models API exposes `capabilities.thinking.types.disabled` (with a `.supported` flag) on `GET /v1/models` and `GET /v1/models/{model_id}`, letting callers check before sending a request whether a model accepts `thinking: {type: "disabled"}` to turn off extended thinking.

— Exact field name and two endpoints given snapshot-20261007

03 Cross-model thinking-block compatibility and refusal behavior

80

BREAKING

Claude Fable 5.1, Claude Opus 5.5, Claude Opus 5, Claude Sonnet 5.5, and Claude Fable 5 may now refuse requests that attempt to elicit internal reasoning as response text, returning an error instead. Thinking-block compatibility across model switches is documented: Claude Opus 5.5 reads thinking blocks from earlier Opus, Sonnet, and Haiku models but not from Fable or Mythos models; Claude Fable 5.1 and Claude Mythos 5.1 read Claude Opus 5.5's blocks on the Claude API; Claude Sonnet 5.5 reads blocks from Claude Sonnet 5, Claude Opus 4.8, Claude Haiku 4.5, and earlier models but not from Claude Opus 5, Claude Opus 5.5, or any Fable or Mythos model. On the Claude API and Google Cloud, switching from Claude Sonnet 5.5 up to Claude Opus 5.5 preserves earlier turns' thinking blocks; any other switch away from Claude Sonnet 5.5, or a switch from Claude Fable 5.1 to Claude Opus 5.5, drops them.

— Detailed rules but no API call to act on product docs

---

04 **Claude Sonnet 4.5 deprecation** DEPRECATED 60

Claude Sonnet 4.5 is deprecated and will be retired on the Claude API on November 30, 2026; users should migrate to an alternative model before that date.

— Exact retirement date given, no migration path detail product docs

└──▷ **BREAKING ON UPGRADE**

! A switch from Claude Fable 5.1 to Claude Opus 5.5 drops earlier turns' thinking blocks.

! Any switch away from Claude Sonnet 5.5 other than to Claude Opus 5.5 (on the Claude API or Google Cloud) drops earlier turns' thinking blocks.

! Claude Sonnet 4.5 is deprecated and will be retired on the Claude API on November 30, 2026; migrate to an alternative model before that date.

WAS THIS USEFUL?  

Google Gemini API

Feedback ⓘ October 6, 2026 NOTES ↗

RANK

Gemini API provides Gemini models for generating text, images, and other multimodal content in applications.

> Gemini API introduced and then graduated its new image model, `gemini-nano-banana-2.1`, to general availability with wide aspect ratio and 4K support, while bundling in several new image-generation capabilities across its Nano Banana model family and deprecating the older Flash Image model.

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

01 **Gemini Nano Banana 2.1 image model reaches GA** NEW

85

`gemini-nano-banana-2.1` is a high-efficiency image generation and conversational editing model running at Flash-level speed and cost, supporting `1K`, `2K`, and `4K` resolutions. It reached general availability with support for aspect ratios `1:4`, `4:1`, `1:8`, and `8:1` at `1K`, `2K`, and `4K` resolutions, along with improved visual quality, prompt adherence, multi-turn character consistency, and text rendering.

— Names model id, aspect ratios and resolutions but no sample call October 6, 2026

02 **New capabilities across Nano Banana image models** NEW

60

Adds Google Image Search Grounding (alongside existing Web Search) for `gemini-nano-banana-2.1` and Gemini 3.1 Flash Image; video-to-image generation support for `gemini-nano-banana-2.1`, Gemini 3.1 Flash Image, and Gemini 3.1 Flash Lite Image; the ability to mix up to 14 reference images in a single workflow on Nano Banana models; controllable thinking levels (default `medium`) for `gemini-nano-banana-2.1`, Gemini 3.1 Flash Image, and Gemini 3.1 Flash Lite Image to balance quality and latency; and character resemblance for up to 4 characters plus fidelity for up to 10 objects in a single workflow on `gemini-nano-banana-2.1`.

— Names models and numeric limits but no parameter or flag syntax product docs

THINNER COVERAGE BELOW

03 **Gemini 3.1 Flash Image deprecated** DEPRECATED

50

The `gemini-3.1-flash-image` model is deprecated; callers should migrate to `gemini-nano-banana-2.1`. No shutdown date has been announced.

— Names both model ids and migration target, no shutdown timeline given October 6, 2026

└──▷ **BREAKING ON UPGRADE**

! The `gemini-3.1-flash-image` model is deprecated; callers should migrate to `gemini-nano-banana-2.1` (no shutdown date announced).

WAS THIS USEFUL? 👍 👎

◆ Open-Weight Models

DeepSeek V4 Pro 0813 Feedback ⓘ

price increase · Oct 6, 2026 MODEL ↗

DeepSeek V4 Pro 0813 is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4-pro-0813): $0.3458 → $1.32 in, $1.0375 → $3.96 out per million tokens (+282%), first seen 6 Oct 2026.

WAS THIS USEFUL? 👍 👎

Z.ai GLM 5.3 Feedback ⓘ price cut · Oct 6, 2026 MODEL ↗

GLM 5.3 is a reasoning model from Z.ai.

Price cut on OpenRouter (z-ai/glm-5.3): $1.40 → $0.07 in, $4.40 → $7.00 out per million tokens (-95%), first seen 6 Oct 2026.

WAS THIS USEFUL? 👍 👎

Google Gemma

Feedback ⓘ October 6, 2026 NOTES ↗ RANK

Gemma is Google's family of downloadable language and multimodal models, including specialist variants.

> Gemma added EmbeddingGemma 2, a new 740M-parameter embedding model.

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

01 **EmbeddingGemma 2 embedding model** NEW 30

Gemma released EmbeddingGemma 2, a new 740M-parameter embedding model.

— Names model and parameter count but no usage detail October 6, 2026

WAS THIS USEFUL? 👍 👎

Meta Llama 3.3 70B Instruct

Feedback ⓘ

price increase · Oct 6, 2026 MODEL ↗

Llama 3.3 70B Instruct is a chat model from Meta.

Price increase on OpenRouter (meta-llama/llama-3.3-70b-instruct): $0.10 → $0.22 in, $0.32 → $0.50 out per million tokens (+120%), first seen 6 Oct 2026.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen3 235B A22B Instruct 2507

Feedback

ⓘ price increase · Oct 6, 2026 MODEL ↗

Qwen3 235B A22B Instruct 2507 is a chat model from Alibaba.

Price increase on OpenRouter (qwen/qwen3-235b-a22b-2507): $0.0875 → $0.09 in, $0.35 → $0.55 out per million tokens (+57%), first seen 6 Oct 2026.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen3.5 397B A17B

Feedback ⓘ

price cut · Oct 6, 2026 MODEL ↗

Qwen3.5 397B A17B is a multimodal model from Alibaba.

Price cut on OpenRouter (qwen/qwen3.5-397b-a17b): $0.55 → $0.45 in, $3.50 → $3.00 out per million tokens (-18%), first seen 6 Oct 2026.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · October 7, 2026

Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue