

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED OCTOBER 23, 2026 · EVERY DAY

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

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

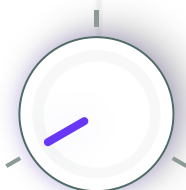
▸ // HOW THIS ISSUE IS MADE

SHOW

VIEW

ISSUE VIEW

scroll



SCROLL

EXPAND
IN
PLACE

PREVIEW
PANE

INFORMATION DENSITY

everything



BRIEFED

EVERYTHING

Every tool in the issue, in full.

all 84 changes

Do you prefer **scroll**?



▸ // CONTENTS SHOW

30 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 shipped a model wave — Voxtral 3 audio, Large 4, Medium 3.5, Small 4 — plus Custom Guardrails
Voice Agents SDK. Bedrock AgentCore Gateway added model

vim ? help

fallback with circuit breaking, so a dead provider stops taking agents down. LangSmith lets you find a past trace by describing it in plain language.

OPERATE

Keep agents running when a model endpoint goes bad

Gateway now falls back to another model and trips a circuit breaker on the failing one, and translates Bedrock InvokeModel wire format across dialects. Retry-and-route logic that teams hand-rolled around provider outages moves into the gateway.

[Amazon Bedrock AgentCore](#)

OPERATE

Find the failing run without writing a trace filter

LangSmith searches runs, traces and threads from a natural-language description, replacing the guess-the-tag-and-metadata-key hunt through the trace UI. Baseten surfaces GPU temperature and utilization on Dedicated Inference deployments, so you can tell a thermally throttled box from a genuinely saturated one.

[LangSmith](#) · [Baseten](#)

BUILD

Hit Cortex and local Macs with the OpenAI SDK you already use

Cortex added an OpenAI-compatible Responses API on its REST API and AI Gateway, so existing Responses clients point at Snowflake with a base-URL change. mlx-serve adds an OpenAI-compatible speech-to-text endpoint with diarisation plus SSD expert-streaming, letting MoE models larger than unified memory run on a Mac.

[Snowflake Cortex AI](#) · [mlx-serve](#)

GOVERN

Write your own content rules instead of taking the vendor's

Custom Guardrails let you define filtering policy against Mistral's models rather than accepting a fixed safety classifier, which is the usual blocker when a domain's normal vocabulary trips a generic filter.

[Mistral AI](#)

OpenAI Codex CLI

Feedback



rust-v0.164.0

NOTES ↗

CODE ↗

RANK

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

Codex CLI's latest release adds preset slots for frequently used prompts, inline arguments for `/model` and `/permissions`, image previews in Kitty/Sixel terminals, enterprise MCP authentication, credential masking for remote sessions, and a default Amazon Bedrock setup wizard.

↳ WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 Numbered preset slots for prompts NEW 60

A new `/presets` command lets users save their most-used prompts and slash commands into numbered slots so they can be recalled instantly during any session.

Save your most-used prompts and slash commands into numbered preset slots so you can recall them instantly during any session.

```
$ /presets
```

– Names command and purpose but no slot limits or storage details rust-v0.164.0

THINNER COVERAGE BELOW

02 Enterprise MCP authentication NEW 55

Adds enterprise MCP authentication using eligible Codex account grants or a dedicated sign-in flow, including support for OAuth clients that require secrets.

– Describes auth mechanism but no setup commands or config rust-v0.164.0

03 Image previews in fullscreen transcript NEW 50

Adds image preview support in the fullscreen transcript for terminals that support Kitty or Sixel graphics protocols.

– Explains what and where, no usage steps given rust-v0.164.0

04 Amazon Bedrock setup wizard by default IMPROVED 50

Enables the Amazon Bedrock setup wizard by default, with API key entry grouped alongside other AWS sign-in methods when credentials are detected.

– Describes behavior but no exact flow or command shown rust-v0.164.0

WAS THIS USEFUL?



All Hands AI OpenHands

Feedback



v1.32.0

NOTES

RANK

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

OpenHands shipped a new MCP install dialog with library selection, cross-provider improvements to its automation picker and pin persistence, and a tightened secret-field policy for self-hosted GitHub connections.

WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 MCP install dialog with library selection NEW 50

Adds an MCP install dialog that lets users choose between installable library options when adding an MCP server.

– Names the dialog but not specific library options or flags. v1.32.0

02 Hidden GitHub token field for self-hosted instances IMPROVED 50

On self-hosted instances where GitHub is already connected, the GitHub token secret field is now hidden in the automations UI, reducing accidental secret exposure.

– Clear behavior and trigger condition but no config/flag name. v1.31.0

03 Automation picker gains cross-provider repos and persistent pins IMPROVED 45

The automation picker now lists every cloneable provider's repositories, enabling cross-provider automation setup, and pinned automations are now stored in agent-server settings so pins persist across sessions.

– Describes two related automation UI changes without exact config keys. v1.32.0

04 Quieter backend canvas logs by default IMPROVED 35

Backend canvas logs are now quieted by default to reduce noise in production deployments.

– Thin description with no configuration detail given. v1.32.0

ALSO FROM THESE RELEASES

Clear macOS Gatekeeper quarantine so the OpenHands Agent Canvas app opens without a **damaged** warning after moving it to Applications.



```
$ xattr -dr com.apple.quarantine "/Applications/OpenHands Agent  
Canvas.app"
```

WAS THIS USEFUL?



GitHub Copilot CLI

Feedback



v1.0.103

NOTES

RANK

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

GitHub Copilot CLI's v1.0.103 release strengthens sandbox handling across ACP sessions, `dnx` commands, and Windows dev-tool shims, retires the legacy `--remote` / `--no-remote` startup flags in favor of clearer launch options, and adds smaller improvements to MCP tool caching, code block copying, session titles, mention suggestions, and GPT-5 verbosity defaults.

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

01 Sandbox handling across ACP, dnx, and Windows shims IMPROVED

83

ACP sessions now honor saved sandbox settings and the `--sandbox` / `--no-sandbox` flags, including when resuming a session (e.g. `copilot --sandbox`); `dnx` commands can reuse existing .NET and NuGet sandbox grants when developer-tool access is enabled; and Windows dev-tool shims like `npm.cmd` now receive the same sandbox grants as `npm`.

Resume a sandboxed ACP session while explicitly enforcing sandbox mode, so saved settings cannot accidentally disable isolation.

```
$ copilot --sandbox
```

– Names flags and a runnable command, but not the full grant-propagation mechanism

v1.0.103

THINNER COVERAGE BELOW

02 Remote launch flags clarified, legacy flags retired BREAKING

50

Remote launch options now use clearer hosting/connection flags, and the legacy `--remote` / `--no-remote` startup flags are retired; any launch scripts using the old flags must be updated to the new flags.

– Names retired flags but not the exact replacement flag names v1.0.103

03 Model-specific default verbosity for GPT-5 IMPROVED

50

Model-specific default Responses verbosity now applies correctly for GPT-5 models unless session verbosity is explicitly set.

– Names GPT-5 models but no flag or config key to act on v1.0.103

04 **Cached MCP tool snapshots shared across Git worktrees**

IMPROVED

45

Copilot CLI now reuses cached local MCP tool snapshots across registered Git worktrees within one repository, avoiding redundant tool discovery.

– Describes behavior but gives no config key or command v1.0.103

05 **Copy original code from completed code blocks**

NEW

35

Clicking a completed response code block copies its original code when enabled.

– Brief description with no setting name or UI path given v1.0.103

06 **Sessions browser shows recorded session titles**

IMPROVED

30

The sessions browser now shows a session's recorded title instead of a generic **Session N** fallback when available.

– Simple UI tweak with no further mechanism detail v1.0.103

07 **Hash mention suggestions refresh with GitHub client**

IMPROVED

30

Hash mention suggestions refresh automatically when the GitHub client becomes ready.

– Minor behavioral fix described without further detail v1.0.103

⚠️ BREAKING ON UPGRADE

! The `--remote` and `--no-remote` startup flags are retired; update any launch scripts to use the new hosting/connection flags instead.

WAS THIS USEFUL?



Cotool

[Feedback](#)[v0.76.0](#)[NOTES ↗](#)

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

Adds email-sending agents, evaluation-score failure alerts, Google Workspace password reset and Microsoft Entra disable/revoke response actions, and improved Proofpoint triage.

Adds a Cotool Email tool so agents can send email to an allowlist of recipients (e.g. replying to employees who reported phishing).

Adds notifications to an agent's failure destinations when a run's evaluation score falls below a minimum threshold you set.

Adds response actions to reset Google Workspace user passwords and to disable Microsoft Entra users and revoke their sessions.

Adds **Created by** and **Updated by** columns to the Agents list.

Improves Proofpoint triage by showing when a threat was removed from mailboxes and flagging verdicts reached after delivery.

NEW INTEGRATION

[Vercel, TruffleHog, Jamf Pro, Material Security, and Zoom integrations](#)

NEW INTEGRATION

[Wiz and Sublime Security managed log ingestion and detections](#)

NEW INTEGRATION

[SEON integration with dedicated webhook trigger](#)

NEW INTEGRATION

[incident.io escalation destination](#)

UPDATED INTEGRATION

[Code42, Databricks, and Wiz connection improvements](#)

UPDATED DETECTION

[Auto-tuning for customized detections](#)

WAS THIS USEFUL?



SST OpenCode

[Feedback](#)[v2.1.3](#)[NOTES](#)[RANK](#)

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

OpenCode shipped new SDK methods for voice model discovery and GPT-Live session hosting, alongside a faster syntax-highlighting engine for Desktop views.

WHAT SHIPPED • 2 FEATURES most completely described first ? what's the number?

01 Voice model discovery and session hosting APIs NEW

80

Adds the `client.voice.models()` SDK method to discover available experimental voice models and supported transports, and `client.voice.session()` to host GPT-Live sessions through connected OpenCode Console workspaces.

Enumerate which experimental voice models and transports your workspace supports before starting a session.

```
javascript
const models = await client.voice.models();
```

Start a GPT-Live voice session through a connected OpenCode Console workspace.

```
javascript
const session = await client.voice.session();
```

– Names both SDK methods with runnable code examples. [v2.1.3](#)

THINNER COVERAGE BELOW

02 Pierre Highlights syntax highlighting engine IMPROVED

50

Switches to Pierre Highlights to reduce highlighting time and memory use for large files and diffs across Desktop code blocks, file views, and review diffs.

– Names the library and affected surfaces but no config or metric. [v2.1.3](#)

WAS THIS USEFUL?



Amazon Kiro

Feedback



CLI: Workflows Start Only When You Ask (2026-10-21)

BLOG

RANK

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

Kiro 2.33.1 tweaks CLI agent management and changes when V3 Workflows run automatically.

WHAT SHIPPED • 2 FEATURES most completely described first ? what's the number?

01 Grant MCP server to specific agent via CLI NEW 80

New `kiro-cli mcp add --agent` command grants a new MCP server to an agent whose tools list did not previously include it, and upgrades the agent file for V3 when `chat.enableAutoAgentUpgrade` is on.

Add an MCP server and grant it to a specific agent that did not previously include it in its tools list.

```
$ kiro-cli mcp add --agent
```

– Exact runnable CLI command with config key named

CLI: Workflows Start Only When You Ask ...

THINNER COVERAGE BELOW

02 Workflows require explicit invocation BREAKING 40

With Workflows enabled in a V3 session, a workflow no longer starts automatically when the feature is turned on — it now only runs when you explicitly ask for one.

– Clear behavior change but no config or flag named

CLI: Workflows Start Only When You Ask ...

BREAKING ON UPGRADE

! With Workflows enabled in a V3 session, a workflow no longer starts automatically when you turn the feature on — it now only runs when you explicitly request one.

WAS THIS USEFUL?



AGENT

Arcade

[Feedback](#)[snapshot-20261023](#)[NOTES](#)[RANK](#)

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

Arcade's single snapshot release adds an opt-in in-cluster usage-reporting collector, expands Helm engine configuration, substantially grows Outlook Calendar and GitHub toolkits, and ships a breaking rebuild of the Google Calendar toolkit alongside several other breaking changes to tool names, output shapes, and default behaviors.

←→ WHAT SHIPPED • 14 FEATURES most completely described first [? what's the number?](#)

01 Engine deployment configuration additions in Helm chart NEW

85

Adds `deployments.resourceQuota.enabled` (set to `false` to disable) with new sizing values when the quota stays on; an optional single connection-string value for the engine's database connection as an alternative to the existing discrete `host` / `port` / `user` fields; and new `workload-identity` values for the engine covering service-account annotations, pod labels and annotations, and extra volumes/mounts.

Switch the engine to a single connection-string instead of managing discrete host/port/user fields — adopt this as its own isolated upgrade.

```
engine:
  database:
    connectionString: 'postgres://user:password@db-
host:5432/arcade'
```

Disable the deployment resource quota when your cluster's policy would otherwise block Arcade workloads.

```
deployments:
  resourceQuota:
    enabled: false
```

— Exact config keys and runnable Helm snippets for two of three additions.

[snapshot-20261023](#)

02 Opt-in in-cluster usage-reporting collector NEW

70

Adds `usageReporting` opt-in config that enables an in-cluster usage-reporting collector, spreading calls across every ready replica for capacity and fault tolerance.

Enable the in-cluster usage-reporting collector in a self-hosted Helm deployment so that usage events are forwarded without a single point of failure.

```
usageReporting:
  enabled: true
```

– Named config key with working example; no numeric limits given. snapshot-20261023

03 Outlook Calendar tool expansion

NEW

70

Expands Outlook Calendar with tools to create, update, delete, and share calendars; tag events with categories; set work location, working hours, and daily meeting hours; set privacy, reminders, hidden guests, Teams links, optional guests, and event time zone; answer and forward every occurrence of a repeating meeting; and read calendar changes since a prior checkpoint.

– Lists many concrete capabilities but no exact tool names or schemas. snapshot-20261023

04 GetFile byte-return now opt-in

BREAKING

60

Changes `GetFile` to return file bytes only when the model's request explicitly asks for them, reducing default response size for large files; callers that relied on bytes being present unconditionally must update their requests to ask for them explicitly.

– Names the exact tool and behavior change but no parameter name shown. snapshot-20261023

THINNER COVERAGE BELOW

05 Worker-readiness check moved out of init container

BREAKING

50

The engine's worker-readiness check moved out of its init container; staged upgrades pinning an older engine image must keep the old init container enabled until the engine image is updated.

– Gives migration note but no flag or config name. snapshot-20261023

06 Google Calendar toolkit rebuilt as new major version

BREAKING

50

Rebuilds the Google Calendar toolkit as a new major version with updated tools and output shapes; callers must migrate to the new major version since the tools and output shapes changed in this breaking rebuild.

– Flags breaking scope but no field-level diff of output shapes. snapshot-20261023

07 Gmail draft composer attachments and HTML editing

IMPROVED

45

Adds Gmail support for attaching files and editing HTML drafts from the draft composer app, not just plain-text drafts.

– Names the composer feature but no tool or endpoint identifier. snapshot-20261023

08 Schema validation for values.yaml

BREAKING

40

`values.yaml` is now schema-validated; a wrongly-typed Helm value that previously succeeded may now fail `helm upgrade`.

– States the mechanism but no migration steps or validation rules. snapshot-20261023

09 Coordinator no longer pulls busybox

BREAKING

40

The coordinator no longer pulls `busybox`; any override or mirror kept solely for that init container should be dropped.

– Names the dropped image but gives no broader context. snapshot-20261023

10 GitHub pull request review tooling

NEW

40

Adds GitHub tools for pull request reviews, changed files, and conversation comments.

– Names three tool areas but no exact tool identifiers. snapshot-20261023

11 Google Contacts expanded search

IMPROVED

35

Expands Google Contacts to search contacts beyond a user's own contacts.

– States the capability only; no tool name or parameter given. snapshot-20261023

12 Tool names changed across worker toolkits

BREAKING

35

Tool names changed across most worker toolkits; a caller using an old tool name now receives an unknown-tool error.

– Flags breaking impact without listing affected toolkits or names. snapshot-20261023

13 Linear tool output shape changed

BREAKING

35

Linear tool outputs changed shape; callers that parse Linear tool responses must update to the new consistent per-entity shape.

– Describes the new shape loosely without field-level detail.

snapshot-20261023

14 Google Calendar self-details tool

NEW

30

Adds a dedicated tool for reading a Google Calendar's own details.

– Thin one-line description with no tool name or schema.

snapshot-20261023

↳ BREAKING ON UPGRADE

- ! Tool names changed across most worker toolkits; a caller using an old tool name now receives an unknown-tool error.
- ! The Google Calendar toolkit's tools and output shapes changed in a breaking rebuild — callers must migrate to the new major version.
- ! Linear tool outputs changed shape; callers that parse Linear tool responses must update to the new consistent per-entity shape.
- ! `values.yaml` is now schema-validated; a wrongly-typed Helm value that previously succeeded may now fail `helm upgrade`.
- ! The engine's worker-readiness check moved out of its init container; staged upgrades pinning an older engine image must keep the old init container enabled until the engine image is updated.
- ! `GetFile` no longer returns file bytes by default; callers that relied on bytes being present unconditionally must update their requests to ask for them explicitly.
- ! The coordinator no longer pulls `busybox`; any override or mirror kept solely for that init container should be dropped.

WAS THIS USEFUL?



Nous Research Hermes

Feedback



v0.22.5

NOTES

CODE

RANK

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

Hermes expanded its terminal backend lineup with serverless, auto-hibernating sandboxes and shipped package-manager pinning controls, new plugin catalog entries, and a Desktop UI improvement for organizing sessions.

WHAT SHIPPED • 3 FEATURES most completely described first ? what's the number?

01 ``pm update --skip-target`` flag NEW 75

Adds the `pm update --skip-target` flag to leave a target's pins untouched during a package-manager update, useful when a plugin's pins must stay locked during a broader upgrade.

Upgrade all packages while preserving a specific target's pinned versions — useful when a plugin's pins must stay locked during a broader update.

```
$ hermes pm update --skip-target my-plugin
```

— Exact flag with a runnable command example v0.22.5

THINNER COVERAGE BELOW

02 Serverless and hibernating terminal backends NEW 57

Hermes adds six built-in terminal backends — local, Docker, SSH, Daytona, Modal, and Vercel Sandbox — plus support for plugin backends such as Singularity. The Daytona and Modal backends offer serverless persistence, hibernating environments when idle to minimize cost.

— Names six backends and hibernation behavior, no setup steps given product docs

03 Drag-to-pin sessions in Desktop sidebar NEW

50

Adds drag-to-pin / unpin sessions in the Desktop sidebar — a user drags a session from the session list and drops it onto the sidebar to pin it, and drags it off to unpin.

Pin or unpin a session in the Desktop sidebar without hunting through menus.

 In the Desktop app, drag a session from the session list and drop it onto the sidebar to pin it; drag it off to unpin.

– Clear UI steps but no deeper mechanism described v0.22.5

WAS THIS USEFUL?



PydanticAI

[Feedback](#)[v2.60.0](#)[NOTES ↗](#)

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

Adds `DockerSandbox`, `BoxLiteSandbox`, `RelevanceCompaction`, `SmartSearch`, and `FileUnderstanding` capabilities; per-run `MCPToolset` auth; new `ToolDefinition` and `RunContext` APIs.

↳ [GET THIS VERSION](#)

```
$ git clone --branch v2.60.0 https://github.com/pydantic/pydantic-ai.git
# already have the repo? check out this version:
$ git checkout v2.60.0
```

Adds `DockerSandbox` harness capability to run agent workspaces inside a local Docker or Podman container.

Adds `BoxLiteSandbox` harness capability for a local BoxLite micro-VM workspace.

Adds `RelevanceCompaction`, which keeps, reduces, or drops older history steps based on a pluggable judge model's ranking.

Adds `SmartSearch` harness capability for plain-English code search with a pluggable judge model.

Adds `FileUnderstanding` harness capability.

+9 more

[NEW SUPPORTED PLATFORM](#) [Bedrock model profiles for xAI Grok 4.6 and 4.7](#)

[NEW INTEGRATION](#) [GitHub Copilot model routing via `github-copilot:` model string](#)

[NEW INTEGRATION](#) [MoonshotAI provider `base\\_url` and `MOONSHOTAI\\_BASE\\_URL` support](#)

[NEW INTEGRATION](#) [ElevenLabs Agents reply attribution via `in\\_response\\_to\\_ids`](#)

[UPDATED INTEGRATION](#) [OpenAIDecisionsModel sends message history as native Responses API input items](#)

[UPDATED INTEGRATION](#) [TemporalDurability supports continue-as-new via `continue as new`](#)

↳ [BREAKING ON UPGRADE](#)

! `RetryPromptPart` is deprecated in favor of `ToolReturnPart(outcome='retried')` and `RetryFeedbackPart`.

- ! `SummarizingCompaction` now summarizes harness-added user-channel content as `<harness_data>` instead of user messages.
- ! `MCP` capability's default `id` is now derived from the server URL's hostname only; two servers on the same host without explicit ids are rejected.
- ! `BedrockConverseModel` now defaults `maxTokens` to Claude's maximum output and streams such requests behind the scenes.
- ! `OpenAIDecisionsProvider` is deprecated; use `OpenAIDecisionsModel` with `OpenAIProvider` directly.
- ! `GitHub` capability now filters tools on `readOnlyHint` under `read_only=True`, rejects `'oauth'`, and deprecates plain `http://` URLs.
- ! Tool returns whose JSON form is a string (e.g. `datetime`, `UUID`) are now sent to the model unquoted so restored histories match the first run.
- ! `sanitize_messages` now strips request-level `instructions` from untrusted history and drops client-supplied `TextContent.metadata` from user prompts.
- ! `SubAgents` background task output is now delivered as framed `<task_output>` data instead of in a `SystemPromptPart`.
- ! `FileSystem` denied and read-only patterns are now matched case-insensitively, `.git` files are protected, and reads and Python regex searches are bounded.

WAS THIS USEFUL?



AutoGPT

[Feedback](#)[autogpt-platform-beta-v0.9.0](#)[NOTES](#)[RANK](#)

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

AutoGPT Platform v0.9.0 ships voice mode and admin abuse-control tools, routes more traffic through an egress proxy, adds budget-aware pause/resume for chats and AutoPilot, and introduces a durable scheduling and DreamPass infrastructure for dream and community jobs.

←→ WHAT SHIPPED · 9 FEATURES most completely described first ? what's the number?

01 Budget-aware pause and resume for chats and AutoPilot

[IMPROVED](#)

74

Adds `send at reset`, an out-of-budget panel for Max, a running-low hint for budget management, and an in-chat Paused card for turns that land at a limit; chats resume or send automatically when the budget resets. AutoPilot turns now pause at their limits instead of terminating, and a `Keep going` action resumes them.

– Names concrete UI elements and resume mechanics, but no API/flag.

[autogpt-platform-beta-v0.9.0](#)

02 Durable scheduling and DreamPass infrastructure for dream jobs

[IMPROVED](#)

65

Adds a durable per-scope schedule registry for dream and community jobs with lifecycle hooks and backfill, plus a durable DreamPass record written from sync and batch dream paths with pass leases, renewal, a reaper for expired passes, retention, and in-flight batch cancellation. Also adds a guard against overlapping dream passes with a stop path built on the DreamPass record.

– Detailed internal mechanism but no user-facing surface to act on.

[autogpt-platform-beta-v0.9.0](#)

03 Admin users page and multi-account abuse tooling

[NEW](#)

60

Adds an admin users page to look up, ban, and impersonate users, with rate limits included, plus additional measures and admin tools against multi-account abuse.

– Names the admin page and its actions; clear navigation starting point.

[autogpt-platform-beta-v0.9.0](#)

THINNER COVERAGE BELOW

- 04

Egress proxy routing for browser tools and Ollama

IMPROVED

55

Routes browser tools' traffic through an egress proxy; traffic from the Ollama block is also routed through the same egress proxy.

– Names the two routed surfaces but gives no config or flag to set it.

autogpt-platform-beta-v0.9.0
- 05

Block-declared effects and risk-based auto-approval

NEW

50

Every block now declares its effect, and Auto judges cheap outward actions automatically while asking first before costly ones.

– Describes the mechanism but no concrete thresholds or flags.

autogpt-platform-beta-v0.9.0
- 06

Voice mode and usage dialogs in chat

NEW

50

Adds voice mode to the  menu in chat; model and token usage now open as dialogs instead of their previous presentation.

– Names the exact UI location but no further mechanism detail.

autogpt-platform-beta-v0.9.0
- 07

Self-hosting via Docker Hub image and Windows installer

NEW

45

Supports self-hosting from the Docker Hub image, with a Windows installer also provided.

– Names the image and installer but no setup steps or commands.

autogpt-platform-beta-v0.9.0
- 08

Per-expert default approval mode for new web chats

NEW

30

Adds a per-expert default approval mode applied to new web chats.

– Bare description with no configuration path or values given.

autogpt-platform-beta-v0.9.0
- 09

Skill distillation during dreaming

NEW

25

Distills private account and Expert skills during dreaming.

– Names the feature but not how distillation works or is used.

autogpt-platform-beta-v0.9.0

WAS THIS USEFUL?



LangChain

Feedback



langchain-core==1.7.0

NOTES >

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

langchain-core 1.7.0 streams tool-call arguments as discrete `tool-call-delta` slices instead of cumulative `block-delta` snapshots.

↳ GET THIS VERSION

```
$ git clone --branch langchain-core==1.7.0  
https://github.com/langchain-ai/langchain.git  
# already have the repo? check out this version:  
$ git checkout langchain-core==1.7.0
```

Streaming `tool-call` arguments now emit `tool-call-delta` events (the new args slice since the last chunk) instead of `block-delta` events (a snapshot of all args so far); `.tool_calls.get()` and `.output` are unchanged.

Iterating `ChatModelStream.tool_calls` now yields each call's new args slice rather than a cumulative snapshot of args received so far.

↳ BREAKING ON UPGRADE

- ! Code that reads raw stream events and inspects `block-delta` args mid-stream must switch to reading `tool-call-delta` events instead; pin `langchain-core<1.7` to keep the old snapshot behaviour.
- ! Iterating `ChatModelStream.tool_calls` now yields the new args slice per chunk, not a cumulative snapshot — code that expected snapshots will see only deltas.

WAS THIS USEFUL?



Agno (formerly Phidata)

[Feedback](#)[v3.1.9](#)[NOTES](#)[RANK](#)

Agno is an agent framework and runtime that orchestrates and runs multi-agent systems.

Agno v3.1.9 adds a direct HTTP endpoint for invoking AgentOS MCP tools without needing an MCP client, along with a way to disable it.

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

01 HTTP endpoint to run AgentOS MCP tools

NEW

90

Adds `POST /mcp/server/tools/{tool_name}/run` to invoke any AgentOS MCP tool directly over HTTP without standing up an MCP client. The new endpoint can be turned off with `MCPConfig(tool_run_endpoint=False)` when it should not be exposed.

Test a specific MCP tool exposed by your AgentOS server without standing up an MCP client — fire an HTTP request directly to the new endpoint.

```
$ curl -X POST
http://localhost:7777/mcp/server/tools/search_web/run \
-H 'Content-Type: application/json' \
-d '{"query": "agno agent framework"}'
```

— Names exact endpoint, config flag, and a runnable curl example

[v3.1.9](#)

WAS THIS USEFUL?



Composio

Feedback



@composio/cli@0.5.1-beta.495

NOTES ↗

RANK

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

Composio's CLI picked up a fail-fast option for dev listeners and richer connection metadata in list output.

↳ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 `--fail-fast` flag for listen commands

NEW

75

Adds a `--fail-fast` flag to the `listen` and `dev listen` CLI commands, stopping execution on the first encountered error — useful in CI pipelines wanting a clean fail signal.

Halt a dev listener immediately on the first error, useful in CI pipelines where you want a clean fail signal.

```
$ composio dev listen --fail-fast
```

— Names exact flag and commands with runnable example @composio/cli@0.5.1-beta.495

THINNER COVERAGE BELOW

02 Connection list shows account ID and alias

IMPROVED

40

The Composio CLI connections list output now includes account ID, alias, and display name for each connection.

— Names fields added but no command or flag given @composio/cli@0.5.1-beta.493

WAS THIS USEFUL?



DEPLOY

CoreWeave

Feedback



snapshot-20261023

NOTES ↗

RANK

CoreWeave is an AI cloud platform offering GPU compute, managed Kubernetes and storage for training and inference.

CoreWeave's SUNK v8.1 release hardens Slurm-on-Kubernetes with per-job workload isolation and a non-root control plane, alongside scheduler annotations, a Slurm version bump, and a mandatory pre-upgrade configuration check.

↳ WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 Per-job workload isolation in SUNK NEW 55

Each Slurm job now gets its own PID namespace, `/tmp`, and `/dev/shm` by default, providing workload isolation between jobs running on shared nodes.

– Names isolation mechanisms but no config or command to invoke snapshot-20261023

02 SUNK scheduler annotations and Slurm version upgrade NEW 50

The SUNK Pod Scheduler now adds `wckey` and `segment` annotations to each job, and SUNK v8.1 upgrades the underlying Slurm version to `25.11.8-coreweave.1`.

– Names annotations and exact version but no usage detail snapshot-20261023

03 Pre-upgrade configuration check required for SUNK v8.1 45

BREAKING

SUNK v8.1.0 requires a configuration check before upgrading — review your setup against the SUNK v8.1 release notes before applying the upgrade.

– Clear actionable warning but lacks specifics of what to check snapshot-20261023

04 Non-root slurmctld on SUNK control plane IMPROVED 40

SUNK now runs `slurmctld` as a non-root user, reducing privilege exposure on the control plane.

– States the change and benefit but it's automatic, nothing to act on snapshot-20261023

↳ BREAKING ON UPGRADE

! SUNK v8.1.0 requires a configuration check before upgrading — review your setup against the SUNK v8.1 release notes before applying the upgrade.

WAS THIS USEFUL?



Eigen Labs Darkbloom

Feedback



v0.10.0

NOTES ↗

CODE ↗

Darkbloom is a for the innovators pushing the frontier.

v0.10.0 adds deeper SSD checkpoint restores on large-memory Macs, staged checkpoint restores, kernel warm-up before serving, and warm MLX buffer pool/paged KV between requests.

↳ GET THIS VERSION

```
$ git clone --branch v0.10.0 https://github.com/Layr-Labs/d-  
inference.git  
# already have the repo? check out this version:  
$ git checkout v0.10.0
```

Defaults to deeper complete-checkpoint SSD restores on large-memory Macs.

Stages SSD checkpoint restores without waiting on another prefill, reducing latency.

Warms first-use kernel shapes before serving to reduce cold-start overhead.

Keeps busy engines' MLX buffer pool and paged KV segments warm between requests.

Applies fused causal SDPA for full-attention prompt chunks.

+2 more

NEW SUPPORTED PLATFORM Nemotron 3.5 Super VL NVFP4

NEW SUPPORTED PLATFORM Nemotron Super video input (mp4/QuickTime).

NEW SUPPORTED PLATFORM Flint model support

NEW SUPPORTED PLATFORM GLM-5.3-Flash serving

NEW INTEGRATION OpenRouter parameter compatibility for Flint

WAS THIS USEFUL?



Baseten

[Feedback](#)[GPU temperature and utilization metrics](#)[NOTES](#)[RANK](#)

Baseten deploys and serves machine learning models as autoscaling inference endpoints, with a model API catalog and dedicated GPU deployments.

Baseten added GPU temperature and utilization metrics to the Metrics tab for Dedicated Inference deployments.

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

01 GPU temperature and utilization metrics NEW 45

The Metrics tab for Dedicated Inference deployments now includes three additional GPU metrics covering temperature and utilization, letting users monitor for thermal throttling or under-utilization directly from the Baseten console's Metrics tab.

Monitor GPU temperature and utilization for a Dedicated Inference deployment to catch thermal throttling or under-utilization.

In the Baseten console, navigate to your Dedicated Inference deployment, then open the Metrics tab to view the new GPU temperature and utilization metrics.

– Names the exact metrics and UI location, but no thresholds or mechanism.

GPU temperature and utilization metrics

WAS THIS USEFUL?



Hugging Face

Feedback



3D Previews for OpenUSD & SimReady Datasets

NOTES >

RANK

Hugging Face is an open platform for building, sharing, and deploying machine-learning models, datasets, and applications.

Hugging Face Hub added interactive 3D previews for OpenUSD files and SimReady datasets, backed by a new open-source viewer package.

↳ WHAT SHIPPED • 1 FEATURE most completely described first ? what's the number?

01 3D previews for OpenUSD and SimReady files on Hub NEW

76

Hugging Face Hub now renders interactive 3D previews of `.usd`, `.usda`, `.usdc`, and `.usdz` files directly on dataset, model, or bucket pages, loading layers, payloads, and textures. Dataset pages containing SimReady packages also gain a **3D Assets** preview section showing package counts aggregated across the dataset, and the rendering is powered by a new open-source `@huggingface/usd-web-viewer` package added to `huggingface.js`.

– Names file formats, package, and UI section; no usage example given.

3D Previews for OpenUSD & SimReady Data...

WAS THIS USEFUL?



Snowflake Cortex AI

Feedback



Oct 23, 2026: Responses API for the Cortex REST API (Preview)

NOTES

RANK

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

Snowflake Cortex AI added an OpenAI-compatible Responses API endpoint to its REST API and AI Gateway, enabling use of the OpenAI SDK or other Responses API-compatible clients against Cortex.

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

01 OpenAI-compatible Responses API endpoint NEW 86

Adds the OpenAI-compatible **Responses API** endpoint at `/api/v2/cortex/v1/responses` on the Cortex REST API and `/v1/responses` on the Cortex AI Gateway, enabling calls to OpenAI models via the OpenAI SDK or any Responses API-compatible client. The API is stateless — the full conversation must be sent with each request — and `previous_response_id` and background mode are not supported.

– Names endpoints and limitations precisely, but feature is in Preview.

Oct 23, 2026: Responses API for the Cor...

WAS THIS USEFUL?



AWS Amazon Bedrock AgentCore

[Feedback](#)

Gateway: Amazon Bedrock InvokeModel support on inference targets (2026-10-22)

[BLOG](#)[RANK](#)

Amazon Bedrock AgentCore provides managed agent runtimes, memory, identity, gateways, observability, and evaluation services.

Amazon Bedrock AgentCore shipped Bedrock InvokeModel wire-format support with cross-dialect translation and model fallback with circuit breaking on Gateway, alongside a breaking v1.0 CLI release with a one-step migration path.

WHAT SHIPPED • 3 FEATURES most completely described first ? what's the number?

01 Bedrock InvokeModel support on Gateway inference targets

[NEW INTEGRATION](#)[NEW](#)

85

Adds support for the `InvokeModel` and `InvokeModelWithResponseStream` Amazon Bedrock API wire format on AgentCore Gateway inference targets, enabling calls in both directions. Also adds a `bedrock-runtime` connector for Gateway inference targets to route through the Bedrock runtime endpoint, with automatic translation between Amazon Bedrock, OpenAI, and Anthropic API dialects.

Call a Gateway inference target using the Bedrock `InvokeModel` wire format when you want existing Bedrock SDK clients to route through AgentCore Gateway without code changes.

```
$ aws bedrock-runtime invoke-model \
  --model-id <inference-target-id> \
  --body '{"prompt": "Hello, world!"}' \
  --endpoint-url https://<agentcore-gateway-endpoint> \
  output.json
```

Call a Gateway inference target with streaming using the `InvokeModelWithResponseStream` API to get token-by-token output through AgentCore Gateway.

```
$ aws bedrock-runtime invoke-model-with-response-stream \
  --model-id <inference-target-id> \
  --body '{"prompt": "Explain quantum computing."}' \
  --endpoint-url https://<agentcore-gateway-endpoint> \
  output_stream.json
```

– Named APIs and connector with runnable invoke commands

02 AgentCore CLI v1.0 with migration and resource commands

85

BREAKING

AgentCore CLI v1 adds an `agentcore migrate` subcommand for upgrading v0 projects to v1, resource commands that operate directly on Runtimes, Memory, Identity, Gateways, and payments with or without a project context, new Runtime templates for scaffolding agent projects, and a redesigned interactive terminal interface. This is a breaking release relative to v0: existing projects must be upgraded via `agentcore migrate`.

Upgrade an existing v0 AgentCore project to the v1 CLI format in one step.

```
$ agentcore migrate
```

– Runnable migrate command plus named resource areas

AgentCore CLI: Version 1.0 is now avail...

03 Model fallback and circuit breaker for Gateway routing **NEW**

70

Adds a `matchModels` condition on gateway rules to match requested models and define an ordered fallback chain in `routeToTarget`, automatically routing to an alternative target or model when the primary destination fails. A service-owned circuit breaker, on by default, skips destinations where requests are failing, requiring no manual intervention to reroute traffic.

– Named config fields explained but no runnable example given

Gateway: model fallback for inference r...

BREAKING ON UPGRADE

! AgentCore CLI v1 includes breaking changes relative to v0 — existing v0 projects must be upgraded using `agentcore migrate`.

Gateway: AWS Step Functions connector (2026-10-21)

BLOG

STABLE

AgentCore Gateway now lets agents start, monitor, stop Step Functions executions and complete task-token callbacks without custom integration code.

WAS THIS USEFUL?



MLX Serve

Feedback



v26.10.3

NOTES ↗

RANK

The mlx-serve runtime provides Apple Silicon inference with OpenAI- and Anthropic-compatible APIs, chat, agent mode, and tool calling.

mlx-serve v26.10.3 adds native Windows builds, two-Mac RDMA clustering over Thunderbolt 5, a new OpenAI-compatible Speech-to-text API with diarisation, Meeting Notes transcription, Neural Engine/Accelerator execution paths, SSD expert-streaming for large MoE models, and several new performance flags.

↳ WHAT SHIPPED · 9 FEATURES most completely described first ? what's the number?

01 New performance and precision CLI flags NEW 85

Adds `--int8-qk` to read very long prompts on dense `Qwen` models 7% faster at 64K tokens and 14% faster at 128K tokens (trading a little precision), plus `--no-lm-head-mx8` to opt out of keeping the output layer in `MXFP8` on a `DGX Spark` and `--lm-head-mx8` to opt in elsewhere.

– Named flag with concrete throughput numbers, no example command. Both flags named verbatim but no usage example given.

v26.10.3

02 Windows native builds via build script NEW 80

Adds `scripts\build-windows.ps1` to build mlx-serve natively on Windows PCs with an NVIDIA graphics card, enabling chat with MLX and GGUF models, vision, images and music; requires the `CUDA` Toolkit.

Build mlx-serve for a Windows NVIDIA workstation from source so you can run MLX and GGUF models on a non-Mac machine.

```
$ scripts\build-windows.ps1
```

– Exact build script path and requirement, fully actionable.

v26.10.3

03 Speech-to-text API with diarisation NEW 80

Adds `Speech-to-text` supporting WAV recordings in 30 languages, with word timings and subtitles in 11 languages and speaker diarisation for up to eight speakers; compatible with any app using OpenAI's transcription API.

– Concrete language/speaker counts, OpenAI-compatible API but no endpoint named.

v26.10.3

04 Two-Mac RDMA clustering over Thunderbolt 5 NEW 75

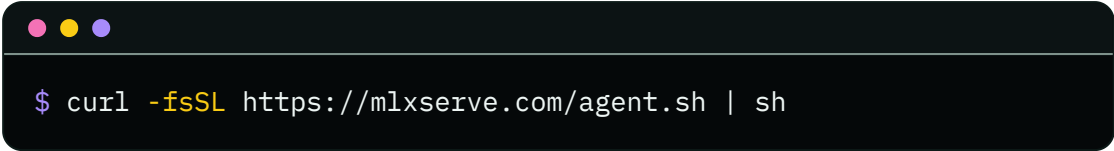
Adds two-Mac RDMA clustering over Thunderbolt 5: pair via Settings > Cluster to run Qwen3.8-Flash-Next and DeepSeek-V4.1-Flash across both Macs, and render MiniMax-H3 video on both simultaneously.

– Exact UI path and named models, no CLI flag. v26.10.3

05 One-line agent configuration script NEW 75

Adds `curl -fsSL https://mlxserve.com/agent.sh | sh` one-liner to configure pi, Claude Code, oh-my-pi, OpenCode or Grok against a local mlx-serve server.

Bootstrap a coding agent (Claude Code, pi, OpenCode, etc.) against your local mlx-serve instance in a single command.



```
$ curl -fsSL https://mlxserve.com/agent.sh | sh
```

– Runnable one-liner command fully specified. v26.10.3

06 Expert-streaming from SSD for large MoE models NEW 70

Adds expert-streaming from SSD for GLM-5.3-Flash and MiMo-V2.6-Flash, enabling them to run on a 128 GB Mac; GLM-5.3-Flash retains its full million-token context at ~35 tokens/second.

– Concrete throughput and memory numbers but no toggle named. v26.10.3

07 Neural Accelerator execution for generative models NEW 65

Adds Neural Accelerator execution for image, video, music and 3D models on M5 Macs with 64 GB or more, running 1.2x–2x faster; toggled via Settings > Neural Accelerator int8.

– Exact UI toggle path and speed multiplier given. v26.10.3

08 Neural Engine execution for dense Qwen models NEW 60

Adds Neural Engine execution for dense Qwen models (Qwen 3 to Qwen3.8), enabled in the model's Settings, freeing the GPU for images, video or another model.

– UI toggle path given, no flag or config key. v26.10.3

THINNER COVERAGE BELOW

09 Meeting Notes via Transcript NEW 50

Adds Meeting Notes via Transcript: records microphone and system audio, then produces a speaker-labeled transcript with summary, decisions and action items from

the active chat model.

– Describes feature but no API, flag or UI path.

v26.10.3

↳ **ALSO FROM THESE RELEASES**

.
.
.

WAS THIS USEFUL?



LocalAI

[Feedback](#)[v4.12.0](#)[NOTES ↗](#)[CODE ↗](#)

LocalAI runs language models locally through self-hosted, OpenAI-compatible APIs.

LocalAI 4.12.0 adds streaming [TTS](#), [/v1/decisions](#) routing, [/v1/images/upscale](#), account-scoped chat history, Qdrant stores, and [MoE inference](#) controls.

↳ [GET THIS VERSION](#)

```
$ git clone --branch v4.12.0 https://github.com/mudler/LocalAI.git
# already have the repo? check out this version:
$ git checkout v4.12.0
```

Adds [POST /v1/decisions](#) endpoint for native decision [model routing](#) with predicates, choices, and scores.

Adds [POST /v1/images/upscale](#) ESRGAN endpoint; Studio records upscaled output as a child image.

Adds [pipeline.streaming.llm](#) / [pipeline.streaming.tts](#) and [pipeline.streaming.clause_chunking](#) config keys for incremental TTS audio delivery.

Adds [moe_cache_mib](#), [lazy_mode](#), and slot [prompt-cache](#) save/restore for workload-specific inference tuning.

Adds [GET /api/conversations](#) for account-scoped server-side chat history requiring authentication and the [chat_history](#) permission.

+2 more

[NEW INTEGRATION](#)[Qdrant persistent vector store backend](#)[NEW INTEGRATION](#)[Remote MCP server OAuth2 client credentials](#)[NEW SUPPORTED PLATFORM](#)[GEM-X motion capture backend on CPU and Vulkan](#)[NEW SUPPORTED PLATFORM](#)[YuE2 music generation support](#)[NEW SUPPORTED PLATFORM](#)[Parakeet bundle multi-role audio support](#)[NEW DETECTION](#)[Responses API PII filtering with protected-terms and adjacent-word masking](#)

WAS THIS USEFUL?



Ollama

[Feedback](#)[v0.40.9](#)[NOTES](#)[CODE](#)[RANK](#)

Ollama downloads, serves, and chats with local language models through a CLI and REST API.

Ollama v0.40.9 adds text-to-speech via Phonon, introduces d3 decision models for classification tasks, extends Flint model support to Linux and Windows, and tightens error handling and context propagation for cloud-backed coding tools.

WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 d3 decision models for classification NEW 90

Adds d3 decision models in four sizes (2B, 4B, 9B, 27B) supporting classification, yes/no answers, and scored probability outputs, invoked via `ollama run d3` with the `--ask` flag.

Classify a customer message against a decision model to route it programmatically without a full chat turn.

```
$ ollama run d3 "A customer says their card was charged twice." -  
-ask "Is this a billing issue?"
```

— Names model sizes, command, and flag with a runnable example. [v0.40.9](#)

02 Phonon text-to-speech synthesis NEW 85

Streams audio through the OpenAI-compatible `/v1/audio/speech` endpoint for text-to-speech generation, and adds `ollama run phonon` CLI playback that plays text or a chat model's reply aloud as audio is generated — both available on Apple Silicon.

Generate spoken audio from text on Apple Silicon — pipe the result or let Ollama play it aloud directly in the terminal.

```
$ ollama run phonon "Hello from Ollama."
```

— Names both endpoint and CLI command with a runnable example. [v0.40.9](#)

03 Error surfacing in Anthropic Messages streaming IMPROVED 65

Streaming `/v1/messages` (Anthropic Messages API) now surfaces in-band runner failures instead of delivering partial output with no error.

— Names the exact API endpoint and before/after behavior, no usage example. [v0.40.9](#)

04 Context propagation for cloud models in launch IMPROVED

60

`ollama launch` now passes cloud models' context lengths and capabilities to coding tools, correcting context and compaction settings for untabled models.

– Names the command and fix but lacks a concrete usage walkthrough. v0.40.9

WAS THIS USEFUL?



EVALUATE

LangChain LangSmith

[Feedback](#)[snapshot-20261023](#)[NOTES](#)[RANK](#)

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

LangSmith's headline addition is AI-powered natural-language search across runs, traces, and threads, paired with new deployment controls (GitHub repo swapping, source provenance) and a broad reshaping of organization permissions (Teams, custom-role Engine permissions, ABAC tabs) that carries several breaking changes to prior permission and deployment-UI behavior.

← WHAT SHIPPED • 9 FEATURES most completely described first ? what's the number?

01 Deployment repo swap and source provenance display

NEW

90

`source_config.repo_url` field added to `PATCH /v2/deployments/{id}` lets you change the GitHub repository of an existing deployment without recreating it; the deployment `Source` section now shows the origin of each deployment — GitHub branch/tag with auto-deploy status, Docker/template image, or CLI (`langgraph deploy`).

Point an existing deployment to a renamed or transferred GitHub repository without recreating the deployment.

```
$ curl -X PATCH
'https://api.smith.langchain.com/v2/deployments/<deployment-id>'
\
  -H 'X-API-Key: <your-api-key>' \
  -H 'Content-Type: application/json' \
  -d '{"source_config": {"repo_url": "https://github.com/my-
org/new-repo-name"}}'
```

— Named endpoint and field with runnable curl example

[snapshot-20261023](#)

02 Thread metadata exclusion in search/count

NEW

85

`exclude_metadata` parameter added to `POST /threads/search` and `POST /threads/count` lets callers filter out threads whose metadata matches any listed object, e.g. automation threads.

Exclude automation threads when searching open threads for a deployed agent, so only real user sessions are returned.

```
$ curl -X POST 'https://api.smith.langchain.com/threads/search' \
-H 'X-API-Key: <your-api-key>' \
-H 'Content-Type: application/json' \
-d '{"exclude_metadata": [{"source": "automation"}]}'
```

– Endpoint, param and runnable curl example all given snapshot-20261023

03 In-pane text search across run data NEW 70

In-pane text search (`Cmd/Ctrl+F`) now searches across run input/output messages, object fields, tags, and metadata, showing a match count with next/previous navigation.

– Exact shortcut and scope named, no further workflow detail snapshot-20261023

04 Org permissions split into per-field controls BREAKING 70

Organization actions previously guarded by a single broad `manage` permission now require dedicated per-field permissions; callers of `PATCH /orgs/current/info` must hold the specific permission for each field they send.

– Named endpoint and before/after permission model snapshot-20261023

05 Header support for tunneled custom MCP servers IMPROVED 65

Custom `MCP` servers behind a LangSmith tunnel now accept headers such as `x-api-key` and `Cookie` without renaming, and OAuth-protected MCP servers can use a tunnel for token, registration, and protected-resource endpoints.

– Named headers and OAuth endpoint types, no setup steps snapshot-20261023

06 ABAC settings page tabs and export IMPROVED 60

The ABAC settings page now includes `Attributes`, `Policies`, and `History` tabs with search, URL-persisted filters, and CSV export.

– Named tabs and capabilities, no exact navigation path given snapshot-20261023

THINNER COVERAGE BELOW

07 CSV export for org members and API keys NEW 55

Organization members and API keys can now be exported to CSV from the Settings page, respecting the current tab, search, and workspace filters.

– Clear capability and scope, but no exact UI control named snapshot-20261023

08 Engine permissions in custom roles NEW 55

Custom roles can now be granted Engine permissions in **Settings > Roles**, with all four Engine permissions available in the role editor.

– Exact settings path given, permission names not enumerated snapshot-20261023

09 Cancellation reason on cancelled runs NEW

40

cancel_reason field added to cancelled runs so callers can read why a run was cancelled.

– Names the field but no example of reading or setting it snapshot-20261023

↳ **BREAKING ON UPGRADE**

- ! Organization actions previously guarded by the single broad **manage** permission now require dedicated per-field permissions; callers of **PATCH /orgs/current/info** must hold the specific permission for each field they send.
- ! Teams APIs return **403** unless the **can_use_teams** setting is enabled for the organization.
- ! The separate Revisions tab on the Deployment page is removed; bookmarks pointing to tabs after Overview now open the tab to their right.
- ! Custom headers for a deployment are no longer shown in the Settings dialog; they are now managed exclusively in the Connect dialog.
- ! Deleting a deployment now requires typing **permanently delete** to confirm.

WAS THIS USEFUL?



Braintrust

[Feedback](#)[snapshot-20261023](#)[NOTES ↗](#)

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

Adds SQL charts, Go and TypeScript SDK capabilities, MCP human-review and dataset-copy flows, Loop on dashboards, coding agent insights, thread view, and bulk prompt/scorer duplication.

Adds SQL chart building for dashboards — write a query or use the visual builder, and Braintrust injects the dashboard's timeframe, filters, and grouping at render time; charts can also be saved from the SQL sandbox (requires data plane v2.1.1+, Pro/Enterprise plans).

Adds Go workflow evaluations (SDK v0.18.0+) allowing Go evals to send tasks and scorers to asynchronous provider APIs (e.g. batch APIs); a run can be resumed by polling, webhook, or from a different process via a Redis store.

Adds TypeScript application logs (SDK v3.39.0+) with severity methods `info()` and `error()` on a logger, and `wrapConsole()` to forward existing console output alongside traces.

NEW INTEGRATION

Extends the Braintrust MCP server to send project log traces, experiment traces, or dataset rows to human review with optional reviewer assignment; Loop automations can queue rows for review on a schedule (requires data plane v2.19.0+).

Extends the Braintrust MCP server and Loop to copy existing project log rows into datasets, preserving input, metadata, tags, and links to source logs (requires data plane v2.18.0+).

[+9 more](#)

NEW SUPPORTED PLATFORM

Cursor session tracing

WAS THIS USEFUL?



AI MODELS

Anthropic

Feedback



snapshot-20261023

NOTES ↗

The Claude Developer Platform is the API for building applications on Anthropic's Claude models that generate, analyze, and automate work.

Fast mode is now available for Claude Opus 5.5 on Claude Models on Google Cloud (Anthropic managed).

NEW INTEGRATION

Claude Models on Google Cloud (Anthropic managed)

WAS THIS USEFUL?



Mistral AI

[Feedback](#)[snapshot-20261023](#)[NOTES](#)[RANK](#)

Learn how to deploy and use Mistral AI's Large Language Models with our comprehensive documentation, guides, and tutorials.

Mistral shipped a major model wave — the Voxtral 3 audio family, Mistral Large 4, Medium 3.5, Small 4, and Z.ai GLM 5.3 Flash — alongside a new Voice Agents SDK, Custom Guardrails, expanded OCR parameters, and inline batching, while also tightening API validation and deprecating several older models.

WHAT SHIPPED • 4 FEATURES most completely described first ? what's the number?

01 OCR API gains granular confidence scores and layout options

NEW

90

Adds `confidence_scores_granularity` supporting `'page'`, `'block'`, and `'word'` granularity (`'page'` returns page-level scores only, `'block'` adds block-level scores, `'word'` adds word-level scores), `include_blocks` (returns a `blocks` array with paragraph-level bounding boxes and structural labels in reading order when set to `true`), `pages` (accepts comma-separated digits/ranges such as `'0,1,2'`, `'0-5'`, or `'0,2-4'`, or a list of integers), `table_format` (choosing between `markdown` and `html`), and `extract_footer` / `extract_header` parameters plus hyperlinks output.

– Names every OCR parameter and its accepted values [snapshot-20261023](#)

02 Model deprecations and rename across OCR, Leanstral, GLM, tiny

DEPRECATED

85

`mistral-tiny` is renamed to `mistral-tiny-2312`, with the old endpoint to be deprecated; `OCR 4.0` (`mistral-ocr-4-0`) is deprecated and retires September 30, 2026 in favor of `mistral-ocr-4-1` or `mistral-ocr-latest`; `Leanstral 1.5` (`labs-leanstral-1-5`) is deprecated and retires September 30, 2026; `Z.ai GLM 5.2` (`zai-glm-5-2`) is deprecated and retires October 31, 2026 in favor of `zai-glm-5-3`.

– Names every deprecated id, retirement date, replacement [snapshot-20261023](#)

03 New model releases: Voxtral 3, Mistral Large/Medium/Small, and Z.ai GLM 5.3 Flash

NEW

80

Releases `voxtral-mini-tts-3-0`, `voxtral-mini-transcribe-3-0`, `voxtral-mini-transcribe-realtime-3-0`, and `voxtral-small-transcribe-3-0` (Voxtral 3 family, Public Preview), spanning TTS with zero-shot voice cloning, streaming and

contextual delivery; offline transcription and diarization with timestamps; and real-time transcription, plus `voxtral-tts-2603` (Voxtral TTS), a text-to-speech model with zero-shot voice cloning, multilingual support, and real-time streaming; `mistral-large-4` (Mistral Large 4), a Public Preview open-weight multimodal Mixture-of-Experts model with a 1M-token context window; `mistral-medium-3-5` (Mistral Medium 3.5), a frontier-class multimodal model with adjustable reasoning via the `reasoning_effort` parameter and open weights under a Modified MIT license; `mistral-small-2603` (Mistral Small 4), a hybrid multimodal model unifying instruct, reasoning, and coding with a 256k context window; and `zai-glm-5-3-flash` (Z.ai GLM 5.3 Flash), a multimodal 320B-parameter (18B active) Mixture-of-Experts model with a 1M-token context window, priced at \$0.15/M input, \$0.03/M cached input, and \$0.50/M output tokens.

– Names every model id and capability, no usage example States spec but no usage guidance Names parameter and license, no example call Names context window, no further mechanism Full spec and pricing, no endpoint usage shown

snapshot-20261023

04 Custom Guardrails for agents and conversations NEW

70

Adds a `guardrails` parameter for configuring Custom Guardrails directly on an Agent, and a `guardrails` field on `POST /v1/conversations` for per-request guardrail configuration.

– Names endpoint and field but not guardrail mechanics snapshot-20261023

↳ BREAKING ON UPGRADE

- ! The new `p` parameter added to Completion API streamed chunks may break applications relying on strict parsing of chunks; users relying on the `mistral-common` package must update to version 1.8.4 or higher (applications using the official SDK are unaffected).
- ! `tool_call_id` is now mandatory in chat messages with the `tool` role; requests omitting it will be rejected.
- ! The API now strictly enforces the validity of all parameters and returns `Extra inputs are not permitted` for unsupported parameters, whereas previously they were silently ignored.
- ! `mistral-tiny-2312` is renamed from `mistral-tiny`; the old endpoint will be deprecated.
- ! `OCR 4.0` (`mistral-ocr-4-0`) is deprecated and retires on September 30, 2026; use `mistral-ocr-4-1` or `mistral-ocr-latest` instead.
- ! `Leanstral 1.5` (`labs-leanstral-1-5`) is deprecated and retires on September 30, 2026.
- ! `Z.ai GLM 5.2` (`zai-glm-5-2`) is deprecated and retires on October 31, 2026; use `zai-glm-5-3` instead.
- ! Documentation paths previously under `/le-chat/*` now redirect to `/vibe/*`; links or integrations using the old `/le-chat/*` or `/mistral-vibe/*` paths will need to be updated.

WAS THIS USEFUL?



◆ Open-Weight Models

MiniMax M3.1 Flash

Feedback



released · Oct 23, 2026

MODEL ↗

MiniMax M3.1 Flash from MiniMax was released on 17 Oct 2026. 1049K-token context; \$0.15 in / \$0.60 out per million tokens (OpenRouter). Open weights on Hugging Face.

WAS THIS USEFUL?



Reflection Beam 1.0 501B A23B

Feedback



price cut · Oct 22, 2026

MODEL ↗

Beam 1.0 501B A23B is a reasoning model from Reflection.

Price cut on OpenRouter (reflection/beam-1.0-501b-a23b): \$0.30 → \$0.25 in, \$1.00 → \$1.00 out per million tokens (-17%), first seen 22 Oct 2026.

WAS THIS USEFUL?



Z.ai GLM 5.3 Flash

Feedback



price cut · Oct 22, 2026

MODEL ↗

GLM 5.3 Flash is a reasoning model from Z.ai.

Price cut on OpenRouter (z-ai/glm-5.3-flash): \$0.048 → \$0.0052 in, \$0.20 → \$0.1985 out per million tokens (-89%), first seen 22 Oct 2026.

WAS THIS USEFUL?



Alibaba Qwen3.5 397B A17B

[Feedback](#)[price increase · Oct 22, 2026](#)[MODEL ↗](#)

Qwen3.5 397B A17B is a multimodal model from Alibaba.

Price increase on OpenRouter (qwen/qwen3.5-397b-a17b): \$0.39 → \$0.50 in, \$2.34 → \$3.60 out per million tokens (+54%), first seen 22 Oct 2026.

WAS THIS USEFUL?



// END OF TRANSMISSION

The AI Toolchain · curated daily · October 23, 2026

Releases & features from the open-source and commercial AI tools that practitioners actually run.

[▶ Subscribe](#)[Feedback on this issue](#)