THE AI TOOLCHAIN

NO. 101

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED NOVEMBER 25, 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

INFORMATION DENSITY

everything



BRIEFED | **EVERYTHING**

Every tool in the issue, in full.

all 72 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 31 of 33 tools

▾ // THE BRIEF HIDE SAME ISSUE, SAME PROMPT, TWO WRITERS: **GPT-5.5** Opus 5

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

Three unrelated things stand out today: Braintrust can trace coding agents across a repo from one config file; Fireworks AI adds cheaper

vim ? help

spot inference for batch/background GLM 5.3 work; SageMaker HyperPod can reshuffle running nodes between instance groups without stopping jobs.

OPERATE

See what a coding agent did across the whole repo without hand-instrumenting each run

A new `.braintrust.json` file and `bt trace` CLI turn repo-wide agent activity into traces. That gives teams a replayable/debuggable record of coding-agent behavior without wiring tracing into every script or relying on chat transcripts after something goes wrong.

Braintrust

DEPLOY

Move lower-priority inference and training work onto cheaper interruptible capacity

Fireworks AI added Spot serverless inference for GLM 5.3 batch and background workloads at half the Standard price, while CoreWeave's SUNK integration now automatically requeues Slurm jobs when Spot Nodes are preempted. The shared gain is practical: teams can use discounted capacity for work that can tolerate interruption without manually babysitting retries.

Fireworks AI · CoreWeave

OPERATE

Rebalance live training infrastructure without killing running workloads

SageMaker HyperPod added an API to move running compute nodes between instance groups without stopping workloads. That makes capacity reshaping a live operations action instead of a drain-and-restart procedure.

Amazon SageMaker AI

GOVERN

Keep untrusted agent extensions from silently expanding what they can execute

Claude Code added a hook to block shell commands loaded by skills, and OpenAI Codex CLI now verifies cryptographic signatures on Plugin Service archives. Both reduce a common agent risk: installing an extension or skill

> no longer has to mean trusting every command or package it tries to bring along.
>
> Claude Code · OpenAI Codex CLI

Does **GPT-5.5** read better? 👍 👎

BUILD

◆ Image & Video Generation

Higgsfield

Feedback ⓘ changelog · Nov 25, 2026 CHANGELOG ↗

Higgsfield is a generative video and image platform for creators, with a developer API.

Higgsfield adds a video-to-video Editing Room, a Property Studio floor-plan-to-tour workflow, and a GPU Harness cloud computer with agent control.

Adds a video-to-video Editing Room with stackable per-moment edits: Swap person, Change angle, Change location, Move camera, Relight, and Edit with brush. Not checked yet

Relight tool supports presets, custom light sources with direction, color, brightness, and diffusion, or a text description. Not checked yet

Review pane lets you run several edits at once and compare each with the original in Split or Compare view before exporting. Not checked yet

Adds Property Studio, a workflow that converts a floor plan into a furnished 3D layout and a narrated video tour with automatic room recognition and labeling. Not checked yet

Property Studio script and voice step supports a library voice, your own voice from a short sample, or a ready recording for narration. Not checked yet

+2 more

NEW INTEGRATION Ideogram 4.6 image model

NEW INTEGRATION Kling Image 4.0 model

NEW INTEGRATION Gemini 4 Neon language model in Supercomputer

NEW INTEGRATION GPT-6.2 Astra and GPT-6.2 Sol in Supercomputer

NEW INTEGRATION Claude Platform direct billing integration: Claude Console

WAS THIS USEFUL? 👍 👎

◆ AI Coding Agents

Augment Code

Feedback ⓘ updated · Nov 25, 2026 NOTES ↗ RANK

Augment Code is an AI coding assistant for VSCode and JetBrains IDEs that provides code review and context-aware suggestions, with Cosmos as its enterprise platform.

> Augment Code's Auggie CLI can now run as an MCP server, letting external MCP-compatible clients query codebase context directly.

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

01 **MCP server mode for Auggie CLI** Not checked yet NEW 83

Run Auggie CLI as an MCP server with `auggie --mcp` to expose the `codebase-retrieval` tool to MCP clients, letting any MCP-compatible client query repo context directly. The `--mcp-auto-workspace` flag automatically selects a workspace per request when running in this mode.

Expose your codebase to any MCP-compatible client so it can query your repo context directly.

```
$ auggie --mcp --mcp-auto-workspace
```

— Names exact flags and tool; runnable command given snapshot-20261125

WAS THIS USEFUL? 👍 👎

Block Goose

Feedback ⓘ v1.61.0 NOTES ↗ CODE ↗ RANK

Goose is an open-source AI agent with a desktop app and CLI for coding and other automated tasks.

> Goose v1.61.0 adds a new `plugin import-server` subcommand for importing MCP Registry server.json files, introduces composable compaction and MCP-extension operations for goose-agent, and ships two new declarative model providers alongside security and workflow tweaks.

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

01 Import MCP server.json via plugin import-server

Not checked yet

NEW

93

A new `plugin import-server` subcommand imports MCP Registry `server.json` files directly into the goose config, with `--package` or `--remote` flags to select a specific zero-based index from the file, or `-` to read the file from stdin.

Bootstrap goose extensions from an MCP Registry server.json you downloaded, selecting the first package entry — useful when onboarding a team to a standard set of MCP tools.

```
$ goose plugin import-server ./server.json --package 0
```

Pipe a server.json from a registry API response directly into goose config without saving the file first.

```
$ curl -fsSL https://registry.example.com/servers/my-tool/server.json | goose plugin import-server -
```

— Runnable commands and named flags given directly v1.61.0

02 Secret-masked headers in mcpServers configs over ACP

Not checked yet IMPROVED

60

All headers in `mcpServers` configs sent over ACP are now treated as secret, preventing credential leakage in logged or surfaced config data.

— Names config key and security mechanism clearly v1.61.0

THINNER COVERAGE BELOW

03 **Composable compaction and MCP extension operations for goose-agent**

Not checked yet NEW 46

goose-agent gains a compaction operation, enabling context-window compaction as a composable agent operation, and an MCP extension operation, making MCP extension management a first-class composable operation.

— Names two operations but no usage shown v1.61.0

04 **New declarative providers: llmman and amazee.ai** Not checked yet

NEW 35

Goose adds llmman as a declarative Ollama-compatible provider, and amazee.ai as a second new declarative provider.

— Providers named but no setup detail given v1.61.0

05 **Optional tool approval for execution** Not checked yet IMPROVED

31

Tool approval is now optional for execution, allowing agent workflows that skip interactive approval prompts.

— No flag or config key named for toggling this v1.61.0

06 **Gemini 4 Neon added to model catalog** Not checked yet NEW

20

Gemini 4 Neon is now available in the model catalog.

— Bare model addition with no further detail v1.61.0

WAS THIS USEFUL?  

Anthropic Claude Code

Feedback ⓘ v2.1.332 NOTES ↗

CODE ↗ RANK

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

> Claude Code v2.1.332 adds a hook to block shell commands loaded by skills, lets MCP servers expose their own skills, and adds a new OpenTelemetry event plus several session-handling and UI notices.

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

01 PreSkillShellCommand hook for skill shell commands

Not checked yet NEW 85

A new `PreSkillShellCommand` hook event can be added to settings to intercept and refuse inline shell commands before a skill or command loads them, letting policy-based hooks block execution at load time.

Block a skill from running any inline shell command during load — useful when you want hook-level control over what gets executed before a skill runs.

json

```
{
  "hooks": {
    "PreSkillShellCommand": [
      {
        "matcher": "*",
        "hooks": [
          {
            "type": "command",
            "command": "echo 'Shell command refused by policy' &&
exit 1"
          }
        ]
      }
    ]
  }
}
```

— Includes exact hook name and a runnable config example v2.1.332

THINNER COVERAGE BELOW

02 MCP servers can expose skills (beta)

Not checked yet NEW 55

MCP servers can now expose skills BETA that appear as `server:skill` in the `/` menu alongside existing skills.

— Names the menu format but lacks setup mechanism v2.1.332

---

03 **OpenTelemetry assistant_response event** Not checked yet NEW

55

A new OpenTelemetry `assistant_response` event captures text Claude sends through `SendUserMessage` or Cowork's message tool when response logging is enabled.

— Names the event and triggers but no config path shown v2.1.332

---

04 **Persistent footer for cross-model flagged messages**

Not checked yet IMPROVED

50

A persistent footer line now stays visible when a flagged message indicates your session is on a different model, remaining until you pick one with `/model`.

— Names the resolving command but little mechanism v2.1.332

---

05 **Code tab session-recovery notices** Not checked yet IMPROVED

50

In VS Code and Desktop's Code tab, a warning now appears when part of a session couldn't be saved, instructing you to run `/compact` to recover it, and a separate notice appears when a resumed session's earlier messages weren't loaded.

— Names the recovery command but not full mechanism v2.1.332

---

06 **Held-message prompt gets 'Always deliver' option** Not checked yet

NEW

30

A new Always deliver option was added to the prompt that appears for held messages from other sessions.

— Bare description of a one-line UI addition v2.1.332

---

07 **Cyber Verification Program tier in startup header** Not checked yet

NEW

25

Your Cyber Verification Program access tier now appears in the startup header beside the model name when signed in with a claude.ai account.

— Thin, display-only change with no further detail v2.1.332

WAS THIS USEFUL? 👍 👎

Factory

Feedback ⓘ

CLI v1.20.0, Desktop v1.20.0: Focus mode in chat, colored file mentions, and fixes acros

BLOG ↗ RANK

Factory's Droids are AI software-development agents with a CLI, web app and API.

> Factory's v1.20.0 release for CLI and Desktop adds self-hosted GitLab merge request support, colored file mentions in chat, and a new focus mode for condensing tool-call activity.

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

01 **Focus mode for chat tool calls** Not checked yet NEW 50

A new focus mode, enabled via a new Appearance setting, folds runs of tool calls into a single summary row in the app.

— Names exact UI setting location, giving a clear starting point.

CLI v1.20.0, Desktop v1.20.0: Focus mod…

02 **Self-hosted GitLab merge request support** Not checked yet NEW 43

Sessions now surface merge requests created on a self-hosted GitLab instance, in addition to existing `GitLab.com` support.

— Names the surface but no setup steps or config key given.

CLI v1.20.0, Desktop v1.20.0: Focus mod…

03 **Colored file mentions in chat** Not checked yet IMPROVED 40

`@file` mentions are now rendered in color so they stand out visually both while typing and in the transcript.

— Names the `@file` syntax but describes only visual change.

CLI v1.20.0, Desktop v1.20.0: Focus mod…

WAS THIS USEFUL? 👍 👎

All Hands AI OpenHands

Feedback ⓘ v1.45.0 NOTES ↗

RANK

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

> This release focuses on UI polish: factories gain worklog visibility and deletion, Canvas gets an experimental Projects view, and the chat sidebar now nests sub-agent conversations under their parents.

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

01 **OpenHands Projects view in Canvas** Not checked yet NEW

50

OpenHands Projects in Canvas is now available as an experimental feature. Project conversations appear in the sidebar with files and pull requests shown beside the chat.

– Describes new experimental UI area and what it shows, no further mechanism. v1.45.0

02 **Factory worklogs and deletion in UI** Not checked yet NEW

45

Factories now show their worklogs newest-first on the factory's own page, and a factory can be deleted directly from that same page in the UI.

– Names the UI surface but no deeper mechanism or limits. v1.45.0

03 **Sub-agent conversation nesting** Not checked yet IMPROVED

35

Sub-agent conversations are now nested under their parent conversation in the frontend.

– Bare statement of UI change with no further detail. v1.45.0

└──▷ **ALSO FROM THESE RELEASES**

Clear macOS Gatekeeper quarantine on the Canvas app after moving it to Applications, so it opens normally without a damaged warning.

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

WAS THIS USEFUL?  

Amazon Kiro

Feedback ⓘ

CLI: Subagent Stop Hooks and Guided Compaction (2026-11-24) BLOG ↗ RANK

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

> Kiro CLI 3.6 introduces SubagentStop hooks, guided `/compact` and `/recap` commands, new settings and a Windows all-users installer, while Kiro separately launches an enterprise referral program and a Kiro Enterprise iPhone app with identity-provider sign-in and web session sync.

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

01 **Enterprise referral program with discount codes** Not checked yet

NEW 83

Enable the referral program for an organization in Controls (off by default); admins can then generate up to 5 referral codes per calendar year under Settings > Referral in Kiro Web. Each code gives the recipient 20% off a Pro, Pro+, Pro Max, or Power plan for 12 billing cycles.

.

As an enterprise admin, turn on the referral program and distribute a discount code to a colleague who wants to subscribe.

📍 1. Go to Controls in Kiro Web and enable the referral program for your organization (off by default). 2. Ask the recipient to open their individual account, go to Settings › Referral, and enter the code. 3. They select Pro, Pro+, Pro Max, or Power — the 20% discount applies for 12 billing cycles starting with their next billing cycle.

— Exact limits, discount percentage, and navigation path given

General: Share Kiro with Friends and Fa…

02 **Kiro Enterprise app for iPhone** Not checked yet NEW 83

Kiro Enterprise for iPhone is now available on the App Store (iOS 26 or later), letting organization account holders start and monitor cloud agent sessions from their phone. Sign-in uses AWS IAM Identity Center or external identity providers such as Okta and Microsoft Entra ID (social login and AWS Builder ID are not supported), and sessions sync between the iPhone app and Kiro Web.

Get started with Kiro Enterprise on iPhone using your organization's identity provider.

📍 1. Download Kiro Enterprise from the App Store (requires iOS 26 or later). 2. Open the app and enter your work email or your organization's start URL. 3.

Complete sign-in through your organization's identity provider (AWS IAM Identity Center, Okta, or Microsoft Entra ID).

— Platform, auth providers, and sync behavior all named

Mobile: Kiro Enterprise for iPhone (202…

03 **Guided /compact summaries in V3 sessions** Not checked yet NEW

80

Add instructions after the `/compact` command in a V3 session to steer what the summary preserves, for example `/compact Keep the migration decisions and remaining validation steps`; the instructions apply only to that compaction.

Compact a long V3 session while preserving specific context — useful before handing off a session or resuming after a break.

```
$ /compact Keep the migration decisions and remaining validation steps
```

— Exact command syntax and scope given with example

CLI: Subagent Stop Hooks and Guided Com…

04 **/recap command and automatic session recap** Not checked yet NEW

78

The new `/recap` command gives a short recap of the current V3 conversation; enabling Session recap triggers an automatic recap after the terminal loses focus for three minutes.

Get a quick summary of where a V3 session stands without scrolling back through the full conversation history.

```
$ /recap
```

— Named command with trigger condition and example

CLI: Subagent Stop Hooks and Guided Com…

05 **Managed settings to lock CLI to V3 sessions** Not checked yet NEW

70

Administrators can set `cli_v2.enabled: false` in the `managed-settings.json` file to prevent managed-device users from launching V2 sessions, locking them to V3.

Prevent managed-device users from launching V2 sessions by locking them to V3 via the managed settings file.

```json
{
  "settings": {
    "cli_v2.enabled": false
  }
}
```

— Exact config key and file shown via example only

CLI: Subagent Stop Hooks and Guided Com…

06 **Notification controls and fullscreen persistence settings**

Not checked yet NEW 65

New notification controls under `/settings` > Notifications let users turn alerts on or off, choose a delivery method, and toggle a tone per alert. The new `/fullscreen` command now saves the fullscreen choice so new sessions start in the last-used mode.

— Named settings path and command but minimal mechanism detail

CLI: Subagent Stop Hooks and Guided Com…

THINNER COVERAGE BELOW

07 **SubagentStop hook for subagent completion** Not checked yet NEW 55

A new `SubagentStop` hook type runs when a dispatched subagent finishes inside a V3 session, letting the hook send one more instruction before the subagent's result returns.

— Names hook type and trigger but no usage example given

CLI: Subagent Stop Hooks and Guided Com…

08 **Large hook output truncation and file save** Not checked yet

IMPROVED 55

Hook output exceeding 3,000 characters is now saved in full to a file; the agent receives the beginning of the output plus the saved file's path, keeping the conversation readable.

— Concrete threshold given but behavior is automatic, no user action

CLI: Subagent Stop Hooks and Guided Com…

09 **Stored login takes precedence over KIRO_API_KEY** Not checked yet

IMPROVED

50

At startup, a stored login now wins over the `KIRO_API_KEY` environment variable, and Kiro CLI reports when it is using the stored login instead of the environment variable.

— Names exact env var and new precedence order CLI: Subagent Stop Hooks and Guided Com…

10 **All-users Windows installer for CLI** Not checked yet NEW

40

Windows stable releases now include an all-users installer so administrators can install Kiro CLI once per machine.

— Single-sentence description with no further mechanism

CLI: Subagent Stop Hooks and Guided Com…

WAS THIS USEFUL? 👍 👎

◆ AI App Builders

Lovable

Feedback ⓘ Nov 23, 2026 (2026-11-24) BLOG ↗

Lovable builds web apps and sites from natural-language prompts.

Lovable adds automated test-user creation for signed-in page testing, a redesigned Security view with per-category pages, Google Docs/Sheets/Slides inline preview in Chats, and auto top-up prompts.

Lovable now creates a temporary test user automatically when testing a page that requires a signed-in user; test users are created only in apps using the built-in backend, require per-project approval (or Always allow) in the project chat, and are removed when Lovable finishes its reply. Not checked yet

The Security view, opened from More → Security, now shows your checks first with Findings and Dependencies on their own pages; select Run checks to run every out-of-date check including advanced checks that read all app code, without consuming credits. Not checked yet

Preview Google Docs, Sheets, and Slides files inline when a Chats message links one — select Preview on the card below the link to open the file in Google's own editor. Not checked yet

Domains now has its own section in the More menu (after Emails) in a project, including in the project menu on mobile. Not checked yet

Lovable now suggests auto top-up above the chat input in a project or in Chats when your workspace buys credits often — select Set up auto top-up to open the auto top-up settings directly. Not checked yet

NEW INTEGRATION Databricks publish integration

NEW INTEGRATION Asaas payment connector

NEW INTEGRATION Paper design tool chat connector

WAS THIS USEFUL? 👍 👎

AGENT

◆ Applied AI Agents

Glean

Feedback ⓘ changelog · Nov 25, 2026 CHANGELOG ↗

Glean is a work AI platform that searches company apps and powers assistants and agents over that data, with an API for developers.

Glean adds real-time Outlook and GitHub agent triggers, Glean skills over MCP, guardrails violation exports via API, an AI Gateway admin section, and artifact comment tools in MCP.

Export guardrails violations by calling the public client API with the `AISECURITY` scope to create, list, inspect, and download export files. Not checked yet

Access Glean skills from any MCP client using the standard `skills/list` and `skills/get` methods, available by default from Glean MCP servers. Not checked yet

NEW INTEGRATION Work with artifact comments and version history from connected AI apps using new tools in Glean's default MCP server. Not checked yet

Restrict Platform API `/api/` requests to IP-allowlisted addresses — the REST API IP allowlist now also applies to `/api/` calls made with API or OAuth tokens. Not checked yet

Agents with Outlook email triggers now start in real time on new messages or threads via webhook notifications instead of polling; new agents use real-time triggers by default and existing eligible agents switch automatically. Not checked yet

+13 more

NEW INTEGRATION Autodesk Construction Cloud connector

NEW INTEGRATION Google Slides MCP server (public beta)

NEW SUPPORTED PLATFORM Claude Fable 5.2 model

NEW SUPPORTED PLATFORM Gemini 4 Carbon model on Vertex AI

NEW SUPPORTED PLATFORM Gemini 4 Hydrogen model

RETIRED SUPPORTED PLATFORM Glean Assist in Zendesk and ServiceNow

RETIRED SUPPORTED PLATFORM Upcoming model deprecations: Gemini 3 Flash, Gemini 3.1 Pro, Gemini 3.1 Pro Custom Tools, GPT Image 2

└──▷ BREAKING ON UPGRADE

! Glean Assist in Zendesk and ServiceNow is retired; support workflows must move to agents in the Glean browser extension.

! The List MCP servers API now returns only the MCP servers a user owns or has a tool available to (unless the user can read all tools); tool audience, credential, and `STDIO`

details are withheld from non-owners unless the caller is an Admin, Setup Admin, or Super Admin.

- ! Reading and editing Salesforce object crawl settings now requires the workspace apps editing permission.
- ! Admin console API responses are now available cross-origin only to trusted, configured client origins — previously credentialed cross-origin access could be granted to any origin.
- ! External search users can no longer call the internal APIs that read or change tool configurations.
- ! Gemini 3 Flash, Gemini 3.1 Pro, and Gemini 3.1 Pro Custom Tools will be deprecated on December 28, 2026; GPT Image 2 will be deprecated on February 6, 2027.

WAS THIS USEFUL?  

◆ Web Search & Scraping for Agents

Perplexity API

Feedback ⓘ updated · Nov 25, 2026 NOTES ↗

Perplexity API provides language models with real-time web search and retrieval for building applications that research and answer questions.

Perplexity API adds the Context API, Monitor API, Batch API, Audit API, Computer MCP incognito mode and usage_get tool, MCP Tasks support, Orchestrator API GA with multi-format requests and auto-routing, and Agent API prompt-cache controls.

New Context API: `POST /v1/context` returns Markdown, HTML, links, relevant passages, and metadata for provided URLs; supports up to five yes/no scoring criteria; client.context.create() in Python 0.48.0 and TypeScript SDK 0.43.0; $1.00 per 1,000 fetched URLs. Not checked yet

New Monitor API: `POST /v1/monitors` creates Search or Page monitors on a schedule with optional webhooks; `client.monitors` in Python 0.47.0 and TypeScript SDK 0.42.0; $0.001 per succeeded run. Not checked yet

New Batch API: submit up to 50,000 Agent API requests as a JSONL file; `client.files` and `client.batches` in the Perplexity Python library and TypeScript SDK; batch requests exempt from Agent API request rate limits. Not checked yet

New Audit API: `GET https://api.perplexity.ai/v1/admin/audit-events` for Enterprise organizations; cursor-based pagination; covers up to 30 days per request. Not checked yet

NEW INTEGRATION Orchestrator API now in general availability: one API key for Anthropic, OpenAI, Google, xAI, and open-weight models; supports OpenAI Chat Completions, OpenAI Responses, and Anthropic Messages formats; `perplexity/auto` for automatic model routing. Not checked yet

+3 more

NEW INTEGRATION Gemini 4 Neon on Orchestrator API: google/gemini-4-neon, google/gemini-4-hydrogen

NEW INTEGRATION GPT-6.2 Sol on Orchestrator API: openai/gpt-6.2-sol, openai/gpt-6.2-astra

NEW SUPPORTED PLATFORM Decider for Mac — local Decisions API on Apple silicon: pplx-decider-flash-v1.2

└──▷ BREAKING ON UPGRADE

! `answer_question`, `confirm_action_approve`, and `confirm_action_deny` now require `followup_uuid` and fail without it.

! `attachment_urls` is no longer supported on `confirm_action_approve` and `confirm_action_deny`.

WAS THIS USEFUL? 👍 👎

◆ AI Agent Frameworks

Arcade

Feedback ⓘ updated · Nov 25, 2026 NOTES ↗

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

Arcade adds multi-account tool profiles, IAM-token PostgreSQL auth, deleted-project data purge, server-side dashboard search/filter/sort/page, and new CLI commands for org suspension and log purging.

Opt organizations into multi-account tool profiles with a multi-profile switch so agents can list, name and switch between authorized accounts for the same tool, with the Engine reporting which profile ran each tool call. Not checked yet

Purge all Engine data for a deleted project — end-user accounts, tokens, secrets, gateways, deployed workers and servers, and tool-execution history — automatically after a configurable retention window; a CLI command handles projects deleted before this release. Not checked yet

Authenticate the Engine, Coordinator and experience API to PostgreSQL using a short-lived IAM token instead of a stored password. Not checked yet

Filter, sort, search and page gateway, secret and OAuth-provider lists from their dashboard pages via server-side operations. Not checked yet

Suspend or unsuspend organizations and accounts from the CLI; remove a suspended organization's deployed workers in one command; credentials under a suspended organization are refused. Not checked yet

+6 more

NEW INTEGRATION Snipe-IT toolkit

UPDATED INTEGRATION Microsoft OneDrive toolkit upgrade

UPDATED INTEGRATION Datadog toolkit metrics querying and charting

└──▷ BREAKING ON UPGRADE

- ! Support for account-bound gateways is removed; the Engine database migration drops the corresponding schema.
- ! The `supportImages.busybox` support image is removed.
- ! The Engine stops serving `Draft` as a protocol version.
- ! The Engine runner image no longer ships `build-essential`, ODBC, `libaio` or GSSAPI runtimes.

WAS THIS USEFUL? 👍 👎

PydanticAI

Feedback ⓘ CVE-2026-136643 CVE ↗ STABLE

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

`CVE-2026-136643`: Online evaluation lets inbound OpenTelemetry baggage set evaluation result attributes (CVSS 5.3, affects `>= 1.85.0, <= 1.107.10`)

Not checked yet

▾ **4 more releases in this issue** · 2026-11-25

> CVE-2026-136642 CVE ↗ STABLE
>
> `CVE-2026-136642`: Realtime sessions on ElevenLabs Agents record credentials from model settings in trace spans (CVSS 6.5, affects `>= 2.58.0, <= 2.73.0`)
>
> Not checked yet

> CVE-2026-136641 CVE ↗ STABLE
>
> `CVE-2026-136641`: `PlaywrightBrowser`'s egress policy does not cover WebRTC, WebTransport, `WebSocketStream` or FedCM traffic (CVSS 6.8, affects `>= 0.24.0, <= 0.73.0`) Not checked yet

> CVE-2026-136640 CVE ↗ STABLE
>
> `CVE-2026-136640`: `PlaywrightBrowser`'s private-address block lets NAT64 forms of cloud metadata addresses and ISATAP forms of private addresses through (CVSS 6.8, affects `>= 0.24.0, <= 0.73.0`) Not checked yet

> CVE-2026-136638 CVE ↗ STABLE
>
> `CVE-2026-136638`: `PlaywrightBrowser` egress policy skips requests whose URL query contains a backslash (CVSS 5.9, affects `>= 0.24.0, <= 0.73.0`)
>
> Not checked yet

WAS THIS USEFUL? 👍 👎

◆ MCP Servers & Tooling

Model Context Protocol Servers

Feedback ⓘ

CVE-2026-134465 CVE ↗

MCP Servers provides reference server implementations for connecting AI applications to external tools and data.

`CVE-2026-134465`: mcp-server-git: nested gitdir file bypasses `--repository` confinement (CVSS 5.3, affects `< 2026.10.28`) Not checked yet

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Inference Providers & GPU Clouds

Cloudflare Workers AI

Feedback ⓘ changelog · Nov 25, 2026

CHANGELOG ↗ RANK

Cloudflare Workers AI runs open-weight models, including the Clef decision models, as serverless inference on Cloudflare's global network.

> Cloudflare shipped Free-plan Email Sending, AVIF image conversion via Polish, expanded image metadata preservation, new feedback channels for CLI and MCP clients, prompt-content logging for AI Security for Apps, password-free two-factor authentication, and removed the legacy `foundation_dns` DNS setting in favor of `nameservers.type`.

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

01 **AVIF image conversion in Polish** Not checked yet NEW 85

AVIF image conversion can be enabled in Polish by setting `avif` to `on` via the REST API endpoint `/zones/{zone_id}/settings/edit/avif`, alongside the existing `polish` and `webp` settings, so browsers that support AVIF receive smaller images than WebP.

Enable AVIF output in Polish so browsers that support AVIF receive smaller images than WebP.

```
$ curl -X PATCH
'https://api.cloudflare.com/client/v4/zones/{zone_id}/settings/ed
it/avif' \
  -H 'Authorization: Bearer <token>' \
  -H 'Content-Type: application/json' \
  -d '{"value":"on"}'
```

— Named endpoint plus runnable curl command changelog-20261125-c3075b2c

02 **Feedback submission via CLI and MCP** Not checked yet NEW 85

Users can send feedback or bug reports using the new `cf feedback` CLI command, optionally authenticated with a `CLOUDFLARE_API_TOKEN` env var, and MCP server clients can submit feedback via the new `submit_feedback` tool, which requires `message`, `type`, and `impact` fields.

File a bug report from the CLI after a confusing deploy output, referencing a submission ID for follow-up.

```
$ CLOUDFLARE_API_TOKEN=<token> cf feedback "The deploy output did not identify the selected environment"
```

– Two named surfaces with a runnable CLI example changelog-20261125-c3075b2c

03 **Removal of foundation_dns DNS setting** Not checked yet BREAKING

85

The `foundation_dns` boolean is removed from the DNS settings endpoints `/zones/{zone_id}/dns_settings` and `/accounts/{account_id}/dns_settings`; PUT requests that include `foundation_dns` will now be rejected, and `nameservers.type: "cloudflare.advanced"` must be used instead.

– Clear migration path but no example request changelog-20261125-c3075b2c

04 **Email Sending available on Free plan** Not checked yet NEW

65

Email Sending is now available on the Free plan, letting a Worker send alert emails to a verified destination address via env.EMAIL.send() without requiring a paid plan.

Send a free alert email from a Worker to a verified destination address during development without a paid plan.

```javascript
export default {
  async fetch(request, env) {
    await env.EMAIL.send({
      from: "alerts@yourdomain.com",
      to: "oncall@example.com",
      subject: "Disk usage above 90%",
      text: "Disk usage on db-1 is above 90%."
    });
    return new Response("Alert sent");
  }
};
```

– Runnable code example but no limits or scope detail changelog-20261125-c3075b2c

05 **Prompt logging for AI Security for Apps** Not checked yet NEW

60

Prompt content can now be logged for AI Security for Apps by configuring payload logging with a WAF custom rule referencing a field with the `cf.llm.prompt` prefix.

— Field prefix named but no full rule example changelog-20261125-c3075b2c

06 **Expanded image metadata preservation** Not checked yet IMPROVED

60

Cloudflare Images now preserves EXIF, XMP, and IPTC-IIM metadata across more image formats using the `metadata` parameter, including EXIF and XMP for PNG, WebP, and AVIF outputs.

— Named parameter and formats but no usage example changelog-20261125-c3075b2c

THINNER COVERAGE BELOW

07 **isVerifiedRecipient field in Email analytics** Not checked yet

NEW

45

Email Sending analytics can now be filtered for verified-recipient messages using the new `isVerifiedRecipient` field in the GraphQL Analytics API.

— Field named but no query example given changelog-20261125-c3075b2c

08 **Password-free two-factor authentication** Not checked yet NEW

45

Two-factor authentication can now be set up without a Cloudflare password by using email verification, available under Profile > Authentication > Two-Factor Authentication.

— UI path given but no mechanism detail changelog-20261125-c3075b2c

└──▷ **BREAKING ON UPGRADE**

! The `foundation_dns` boolean is removed from DNS settings endpoints `/zones/{zone_id}/dns_settings` and `/accounts/{account_id}/dns_settings`; PUT requests that include `foundation_dns` will be rejected — use `nameservers.type: "cloudflare.advanced"` instead.

WAS THIS USEFUL? 👍 👎

CoreWeave

Feedback ⓘ updated · Nov 25, 2026 NOTES ↗ RANK

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

> CoreWeave's SUNK Slurm integration gained automatic requeueing on Spot Node preemption and changed how user provisioning matches groups, the latter being a breaking change.

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

01 **SUNK user provisioning by POSIX group name** Not checked yet

60

BREAKING

SUNK v8.3.0 changes user provisioning to select users by each group's POSIX group name instead of its display name; operators should check their group configuration before upgrading.

— Names exact version and migration check needed snapshot-20261125

THINNER COVERAGE BELOW

02 **Automatic Slurm requeue on Spot preemption** Not checked yet NEW

40

SUNK now requeues Slurm jobs automatically when a Spot Node receives a preemption warning, reducing the need for manual intervention during preemptions.

— Describes behaviour but no config or command to invoke it snapshot-20261125

└──▷ **BREAKING ON UPGRADE**

! SUNK v8.3.0 changes user provisioning to select users by POSIX group name instead of display name — check your group configuration before upgrading.

WAS THIS USEFUL? 👍 👎

Fireworks AI

Feedback ⓘ New Serverless mode: Spot NOTES ↗

RANK

Fireworks AI provides hosted inference, model fine-tuning, and deployment APIs for open-weight models.

> Fireworks AI introduced a Spot serverless inference mode for GLM 5.3, offering batch and background workloads at half the Standard price.

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

01 **Spot serverless inference mode** Not checked yet NEW 90

A new Spot serverless mode runs inference for batch-style and background workloads at 50% of the Standard price. Enable it by setting `service_tier: "spot"` in API requests with model `accounts/fireworks/models/glm-5p3`, or by pointing `model` directly to `accounts/fireworks/models/glm-5p3-spot`. Spot is available for GLM 5.3, priced at $0.70 input / $0.13 cached input / $2.20 output per 1M tokens, for Tier 3+ and Enterprise accounts.

Run a large-scale eval or data-generation job against GLM 5.3 at half the cost by routing it to the Spot tier via the API.

```
$ curl https://api.fireworks.ai/inference/v1/chat/completions \
  -H 'Authorization: Bearer <API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"model": "accounts/fireworks/models/glm-5p3",
"service_tier": "spot", "messages": [{"role": "user", "content":
"Summarise this document: <text>"}]}'
```

Use the dedicated Spot model alias to send background jobs to Spot tier without specifying `service_tier` on every request.

```
$ curl https://api.fireworks.ai/inference/v1/chat/completions \
  -H 'Authorization: Bearer <API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"model": "accounts/fireworks/models/glm-5p3-spot",
"messages": [{"role": "user", "content": "Generate 100 training
examples for <task>"}]}'
```

— Names exact flag, model IDs, pricing and account tiers with runnable examples.

New Serverless mode: Spot

WAS THIS USEFUL?

AWS Amazon SageMaker AI

Feedback ⓘ November 20, 2026

NOTES ↗ RANK

Amazon SageMaker AI provides managed tools and infrastructure for building, training, fine-tuning, and deploying machine-learning models.

> SageMaker HyperPod adds a new API to move running compute nodes between instance groups without stopping workloads.

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

01 **MoveClusterNodes API for HyperPod instance groups** Not checked yet

NEW 82

The new `MoveClusterNodes` API lets you move running nodes between instance groups within a cluster or between clusters in the same AWS account and Region, for example rebalancing a node from a development group to a production group without stopping workloads. It can be invoked via `aws sagemaker move-cluster-nodes --source-cluster-name <source-cluster> --target-cluster-name <target-cluster> --node-ids <node-id> --target-instance-group-name <target-group>`.

Rebalance compute by moving a running node to a different instance group — for example, shifting a node from a development group to a production group without stopping workloads.

```
$ aws sagemaker move-cluster-nodes --source-cluster-name <source-cluster> --target-cluster-name <target-cluster> --node-ids <node-id> --target-instance-group-name <target-group>
```

— Names exact API and runnable CLI command with parameters November 20, 2026

WAS THIS USEFUL?

◆ ML Platforms & SDKs

Palantir AIP

Feedback ⓘ

GPT-6.2 Astra and GPT-6.2 Sol now available in AIP NOTES ↗ STABLE

Palantir's Artificial Intelligence Platform (AIP) connects AI with your data and operations, with builder tools for production AI-powered workflows, agents, and functions on top of the Ontology.

AIP adds GPT-6.2 Astra for complex multi-step projects and GPT-6.2 Sol for everyday coding and agent workloads, with geo-restricted enrollment support.

NEW INTEGRATION Access GPT-6.2 Astra through Amazon Bedrock, Azure OpenAI, or Direct OpenAI in AIP for multi-step complex tasks such as revising software across several steps or analyzing sets of documents. Not checked yet

Access GPT-6.2 Sol through Amazon Bedrock, Azure OpenAI, or Direct OpenAI in AIP for everyday work including coding tasks and running agents at scale. Not checked yet

NEW SUPPORTED PLATFORM US-georestricted commercial enrollments can access GPT-6.2 Astra and GPT-6.2 Sol through Direct OpenAI only. Not checked yet

NEW INTEGRATION Gemini 4 Neon via Google Vertex AI

NEW SUPPORTED PLATFORM US government infrastructure support for Gemini 4 Neon (IL2, IL4, IL5): FedRAMP High, Defense Industrial Base, Department of Defense

▾ 2 more releases in this issue · 2026-11-24

> Gemini 4 Argon now available in AIP on UK-georestricted enrollments NOTES ↗ STABLE

> Gemini 4 Neon now available in AIP through Google Vertex AI NOTES ↗ LATEST STABLE
>
> Gemini 4 Neon adds tool calling and structured JSON output matching a caller-defined schema, with a 1,048,576-token context window shared across input and output.
>
> Gemini 4 Neon supports tool calling and structured JSON output matching a caller-defined schema. Not checked yet
>
> AIP now supports Gemini 4 Neon via Google Vertex AI for eligible commercial and US government enrollments, accepting text, images, and audio with a 1,048,576-token context window shared across input and output. Not checked yet

WAS THIS USEFUL? 👍 👎

◆ AI Chat Apps

Anthropic Claude App

Feedback ⓘ November 24, 2026

NOTES ↗ RANK

Claude App is Anthropic's Claude assistant on web, desktop and mobile, where chat and Cowork now run as one unified Claude experience.

> Claude App's biggest shift this window is bringing computer use and Google Meet participation into beta, alongside installable artifact apps, file transfers to connected tools, and a generally-available monthly recap.

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

01 Monthly recap (Reflect) reaches general availability

Not checked yet IMPROVED 80

Monthly recap, found in Settings under Reflect, is now out of beta on Free, Pro, Max, and Ultra plans across web, Claude Desktop, iOS, and Android. Recaps can now draw on Claude Code sessions and Claude Cowork tasks, and stored recaps can be deleted from the Reflect page.

Review or delete your stored monthly recaps after the feature moves out of beta.

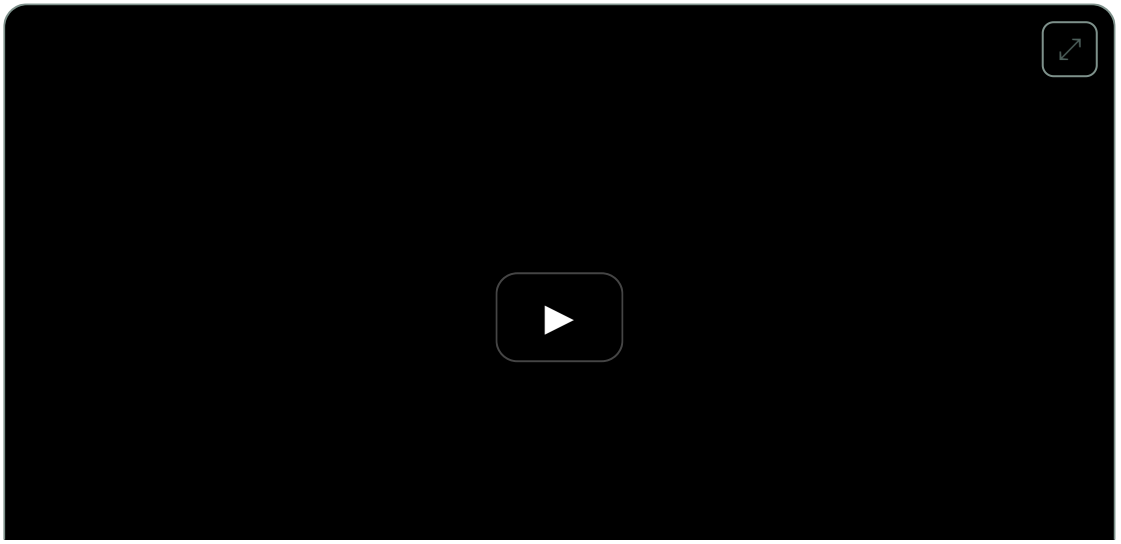
📍 Go to Settings › Reflect to view your monthly recap history. Use the delete option on the page to remove any stored recaps.

— Exact settings path and delete action given, plus source data named November 24, 2026

02 Release channel setting in Claude Desktop

Not checked yet NEW 75

A new release channel setting lets users choose whether Claude Desktop updates on a stable (less frequent, more tested) or latest (daily) track, via Settings > Release Channel.





Opt into daily Claude Desktop updates to get fixes and new features as soon as they ship

> 📍 In Claude Desktop, go to Settings › Release Channel and select `Latest` to receive updates approximately once a day.

– Exact setting path and both track options are given November 17, 2026

---

03 **Claude Meetings joins Google Meet calls** Not checked yet NEW

75

Claude Meetings BETA, available on Team and Enterprise plans, lets Claude join a Google Meet call as a visible participant to take notes, answer questions, and build artifacts such as docs, whiteboards, and slides during the call.

Add Claude to a live Google Meet call on a Team or Enterprise plan to get real-time notes and artifact generation without leaving the meeting.

> 📍 In Claude or Claude Desktop, open Claude Meetings BETA and add Claude as a participant to your active Google Meet call.

– Describes behavior and plans but entry point in UI is vague November 12, 2026

---

04 **Install Claude-generated artifacts as native desktop apps**

Not checked yet NEW

70

Claude can turn an artifact it builds into an installable native app on Mac or Windows BETA; the app gets its own window and icon and can interact with the operating system like any other app. Users install it via the artifact menu's `Install as app` option.

Turn a Claude-generated artifact into a standalone desktop app you can launch from your dock or taskbar.

> 📍 In the artifact view, open the artifact menu and select `Install as app`. The app will be installed on your Mac or Windows computer with its own window and icon.

– Names exact menu action but mechanism of OS interaction stays vague November 24, 2026

---

05 **File transfers to and from connected apps** Not checked yet NEW

70

Claude can send files to connected apps and save files returned by them directly inside a conversation (beta, rolling out gradually) on Pro, Max, Ultra, Team, and Enterprise plans; Enterprise requires an owner to turn it on.

Save a file returned by a connected app directly to your computer during a Claude conversation.

> 📍 With a supported app connected, ask Claude to retrieve or generate a file through that app. When the app offers a file, Claude will present a `Save file` option inline in the conversation.

— Plans and rollout named but trigger flow described loosely November 24, 2026

---

06 **Computer use for Team and Enterprise** Not checked yet NEW

65

Computer use BETA is now available on Team and Enterprise plans, letting Claude see a user's screen and act in desktop apps the user allows, across Claude Desktop, Claude on the web, and Claude Mobile wherever `Computer access` is turned on.

— Names plans and the access toggle but no exact navigation path November 20, 2026

THINNER COVERAGE BELOW

07 **Refreshed mobile look for iOS 26** Not checked yet IMPROVED

25

A refreshed mobile look is rolling out for Claude on iOS, requiring iOS 26 or later; no opt-in is needed, users just need to keep the app updated.

— No detail on what the redesign actually changes November 20, 2026

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

Mintplex Labs AnythingLLM Feedback ⓘ CVE-2026-55611

CVE ↗

AnythingLLM is a local-first desktop and server application that runs agents, RAG workspaces, and MCP tools against local or hosted models.

`CVE-2026-55611`: Incomplete fix for GHSA-p5rf-8p88-979c: embed-parsed-file cleanup deletes any parsed file by ID without ownership scoping (cross-tenant IDOR deletion) (affects `all versions carrying the GHSA-p5rf-8p88-979c fix through the current main`) Not checked yet

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

Braintrust

Feedback ⓘ updated · Nov 25, 2026 NOTES ↗ RANK

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

> Braintrust shipped repo-wide coding-agent tracing via a new `.braintrust.json` file and `bt trace` CLI, a Decisions API for LLM-as-a-judge scorers and classifiers, scheduled project-activity digests, and a new Gateway token-usage page, alongside webhook header support, a `--service-tier` flag for scorer creation, and Terraform provider updates including a Patterns view resource and a raised minimum Terraform/OpenTofu version.

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

01 Repo-wide coding-agent tracing via bt trace

Not checked yet NEW

90

A `.braintrust.json` file at the Git repo root configures tracing for all supported coding agents in that repository, installed and managed with `bt trace install` (for example `bt trace install --scope project` to trace every agent run in the codebase to one project). `bt trace run --inherit` starts a coding-agent session using the saved repository settings.

Enable repository-wide tracing for all supported coding agents in a repo, so every agent run in this codebase is traced to the same project without per-agent config.

```
$ bt trace install --scope project
```

— Names config file and exact commands with a runnable example snapshot-20261125

02 bt trace disable now pauses instead of removing

Not checked yet

75

BREAKING

`bt trace disable` now pauses tracing while keeping the integration and its settings, instead of removing them entirely as before; scripts that previously relied on `bt trace disable` to remove tracing should switch to `bt trace uninstall`.

— Clear before/after with a named migration command snapshot-20261125

03 --service-tier flag for scorer creation

Not checked yet NEW

70

`bt scorers create` now accepts a `--service-tier` flag to set the OpenAI service tier (for example, `flex`) used when creating scorers.

— Exact flag and example value, no full command example snapshot-20261125

04 **Terraform provider Patterns view resource** Not checked yet NEW

70

The Terraform provider's `braintrust_view` resource can manage a Patterns saved view as code (`view_type = "patterns"`), letting it be version-controlled and deployed alongside other infrastructure.

Manage a Patterns saved view as Terraform code so it can be version-controlled and deployed alongside other infrastructure.

hcl

```hcl
resource "braintrust_view" "patterns_overview" {
  view_type = "patterns"
  name      = "Patterns overview"
}
```

— Named resource type with runnable HCL example snapshot-20261125

05 **Decisions API for scorers and classifiers** Not checked yet NEW

60

The Decisions API lets LLM-as-a-judge scorers and classifiers have the model select one of your defined choices directly; each result records a confidence score and per-choice probability to surface uncertain judgments.

— Explains mechanism but no endpoint or syntax given snapshot-20261125

06 **Gateway token usage in Inference activity page** Not checked yet

NEW

60

A new `Settings > Inference activity` page shows Braintrust Gateway token usage, filterable by project, feature, and whether requests use built-in models or your own provider keys.

— Exact UI path with filters named snapshot-20261125

THINNER COVERAGE BELOW

07 **Custom headers for webhook destinations** Not checked yet NEW

55

Webhook destinations used for alerts, Patterns, and Loop automations can now include custom HTTP headers, including encrypted credential values.

— Names the three surfaces but no config syntax given snapshot-20261125

08 **Raised minimum Terraform/OpenTofu version** Not checked yet

50

BREAKING

The Terraform provider now requires Terraform or OpenTofu 1.11.1 or later; previously lower versions were supported.

— Exact version threshold but no migration steps snapshot-20261125

09 **Scheduled project-activity digests** Not checked yet NEW

45

Scheduled digests send recurring project-activity summaries to your team, with a choice of compact or detailed writing and configurable destinations.

— Describes options but no config key or setup steps snapshot-20261125

└──▷ **BREAKING ON UPGRADE**

! `bt trace disable` now pauses tracing and keeps the integration and its settings instead of removing them; use `bt trace uninstall` in scripts that previously used `bt trace disable` to remove tracing.

! The Terraform provider now requires Terraform or OpenTofu 1.11.1 or later (previously lower versions were supported).

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI Security Posture & Discovery

Mend.io Mend AI

Feedback ⓘ changelog · Nov 25, 2026

CHANGELOG ↗ RANK

Mend AI is an AI security service that discovers AI components and helps teams govern their use in software applications.

> Mend AI shipped a scan history view for tracking AI risk trends over time and added autocomplete to AI Analysis workflow provider fields.

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

01 **Autocomplete for AI Analysis workflow provider fields**

Not checked yet IMPROVED 60

AI Analysis workflows now include autocomplete with a free-text fallback for the Model Provider and Model Inference Provider author fields, allowing any detected vendor or a manually entered value to be selected when creating workflow rules.

— Names exact fields but gives no exact UI path or endpoint changelog-20261125-889150d5

THINNER COVERAGE BELOW

02 **AI scan history with trend tracking** Not checked yet NEW

45

A new AI scan history view lets users track AI risk trends over time, compare findings across scans, and drill down into individual scan results.

Review AI risk trends across branches over time to spot regressions in a project's AI posture.

> 📍 In the Mend AI platform, navigate to AI Scan History, select a repository and branch, then compare findings across scans and drill down into individual scan results.

— Describes behaviour but no concrete surface beyond UI navigation

changelog-20261125-889150d5

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Open-Weight Models

Xiaomi MiMo-V2.6-Flash-MOPD

Feedback ⓘ

released · Nov 25, 2026 MODEL ↗

MiMo-V2.6-Flash-MOPD from Xiaomi was released on 23 Nov 2026. 1049K-token context; $0.50 in / $1.00 out per million tokens (OpenRouter). Not checked yet

WAS THIS USEFUL? 👍 👎

DeepSeek V4 Pro 0813

Feedback ⓘ

price increase · Nov 24, 2026 MODEL ↗

DeepSeek V4 Pro 0813 is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4-pro): $0.90 → $1.28 in, $1.50 → $2.56 out per million tokens (+71%), first seen 24 Nov 2026. Not checked yet

WAS THIS USEFUL? 👍 👎

Tencent Hy3

Feedback ⓘ price cut · Nov 24, 2026 MODEL ↗

Hy3 is a chat model from Tencent.

Price cut on OpenRouter (tencent/hy3): $0.14 → $0.132 in, $0.58 → $0.528 out per million tokens (-9%), first seen 24 Nov 2026. Not checked yet

WAS THIS USEFUL? 👍 👎

Thinking Machines Inkling Feedback ⓘ

context change · Nov 24, 2026 MODEL ↗

Inkling is a multimodal model from Thinking Machines.

Context window grew from 524K to 1049K tokens on OpenRouter (thinkingmachines/inkling), first seen 24 Nov 2026. Not checked yet

WAS THIS USEFUL? 👍 👎

inclusionAI Ling 3.1 Flash Feedback ⓘ

context change · Nov 24, 2026 MODEL ↗ STABLE

Ling 3.1 Flash is a chat model from inclusionAI.

Context window grew from 1M to 1049K tokens on OpenRouter (inclusionai/ling-3.1-flash), first seen 24 Nov 2026. Not checked yet

▾ 1 more release in this issue · 2026-11-24

> price increase · Nov 24, 2026 MODEL ↗ STABLE
>
> Price increase on OpenRouter (inclusionai/ling-3.1-flash): $0.06 → $0.075 in, $0.18 → $0.225 out per million tokens (+25%), first seen 24 Nov 2026. Not checked yet

WAS THIS USEFUL? 👍 👎

PrismML

Feedback ⓘ

Introducing Ternary Bonsai 2.5 27B: Powering Agentic Coding and Private Knowledge Work (

BLOG ↗ RANK

PrismML builds 1-bit and ternary Bonsai models that pack advanced reasoning, tool use, and multimodal AI into phones, laptops, and efficient data centers.

> PrismML shipped Bonsai 2.5, a new ternary-quantised flagship model line topped by a 27B model that shrinks to under 6 GB while matching FP16 on tool-calling, and swapped it in as the default demo model.

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

01 **Bonsai 2.5 27B ternary quantisation and performance**

Not checked yet NEW 87

Bonsai 2.5 27B now fits in `5.93 GB` (down from ~54 GB at FP16), via sub-two-bit-per-weight quantisation that reduces memory-bandwidth pressure, enabling agentic coding and tool-calling on a laptop or a single GPU while matching FP16 on agentic tool calling and retaining 90% of its agentic coding score. It decodes at approximately 160 tok/s on an RTX 5090, approximately 300 tok/s with its draft head, and approximately 30 tok/s on an M5 Pro laptop, and ships as a GGUF pack runnable on upstream llama.cpp with a separate `Q8_0` vision projector (0.36–0.63 GB) for image input.

– Names format, quant scheme and benchmark numbers; missing exact run command

Introducing Ternary Bonsai 2.5 27B: Pow…

THINNER COVERAGE BELOW

02 **Bonsai 2.5 flagship model line** Not checked yet NEW 41

PrismML adds Bonsai 2.5 as the new flagship model line, built for agentic coding and long, tool-driven tasks, available in four sizes from 2B to 27B, with the 27B variant offering stronger agentic coding, multi-turn tool use, and multi-step task execution.

– Names sizes and purpose but no mechanism or usage path

Introducing Ternary Bonsai 2.5 27B: Pow…

03 **Bonsai 2.5 27B replaces prior default demo model** Not checked yet 40

BREAKING

Bonsai 2.5 27B replaces Ternary Bonsai 2 27B as the default model in the demo repository; prior weights remain available on Hugging Face.

– States the change and fallback location, no migration steps product docs

└──▷ **BREAKING ON UPGRADE**

! Bonsai 2.5 27B replaces Ternary Bonsai 2 27B as the default model in the demo repository; prior weights remain on Hugging Face.

WAS THIS USEFUL? 👍 👎

StepFun Step 3.7 Flash Feedback ⓘ

context change · Nov 24, 2026 MODEL ↗

Step 3.7 Flash is a reasoning model from StepFun.

Context window shrank from 262K to 256K tokens on OpenRouter (stepfun/step-3.7-flash), first seen 24 Nov 2026. Not checked yet

WAS THIS USEFUL? 👍 👎

◆ **Frontier Models**

Upstage Solar Mini 4.5 (Preview) Feedback ⓘ

released · Nov 25, 2026 MODEL ↗

Solar Mini 4.5 PREVIEW from Upstage was released on 24 Nov 2026. 524K-token context; $0.01 in / $0.04 out per million tokens (OpenRouter). Not checked yet

WAS THIS USEFUL? 👍 👎

Upstage Solar Pro 4 Contributor Feedback ⓘ

released · Nov 25, 2026 MODEL ↗

Solar Pro 4 Contributor from Upstage was released on 24 Nov 2026. 524K-token context; $0.09 in / $0.36 out per million tokens (OpenRouter). Not checked yet

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · November 25, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue